AN INTRODUCTION TO

MOLECULAR ANTHROPOLOGY

MARK STONEKING

WILEY Blackwell

AN INTRODUCTION TO MOLECULAR ANTHROPOLOGY

AN INTRODUCTION TO MOLECULAR ANTHROPOLOGY

Mark Stoneking

Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany

WILEY Blackwell

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Stoneking, Mark.
An introduction to molecular anthropology / Mark Stoneking.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-06162-6 (pbk. : alk. paper) 1. Human genetics–Variation. 2. Human molecular genetics. 3. Molecular evolution. 4. Human evolution. I. Title.
QH431.S784 2015
572.8–dc23

2014039320

10 9 8 7 6 5 4 3 2 1

For my students, who I hope have learned as much from me as I have from them.

"A study of the blood of individual nations enables us to decode their distant past"

—Ludwik Hirszfeld, *A Story of One Life*, 1946

"Whether or not it is true that the proper study of mankind is man, it is certain that he finds great difficulty in studying anything else."

—John William Navin Sullivan, *Aspects of Science*, 1923

CONTENTS

PREFACE

When most people think about anthropology, the image that usually comes to mind is that of intrepid, Indiana Jones-like characters, traveling to remote and exotic locations; living and working under arduous conditions; digging up fossils, stone tools, or other evidence of our past; and making headlines by proclaiming that what they have found overturns everything we thought we knew about human evolution. However, there is another type of anthropology that is becoming an increasingly important source of information about our past, rivaling the study of fossils or artifacts, and that is **molecular anthropology**, which can be defined as the use of molecular genetic methods to address questions and issues of anthropological interest. More specifically, molecular anthropology uses genetic evidence to obtain insights into human origins, migrations, and population history, as well as the role of natural selection during human evolution, and the impact of particular cultural practices on patterns of human genetic variation. And while working in a molecular genetics laboratory or sitting in front of a computer (which is where most of the work is done nowadays) may lack the glamour and excitement of paleoanthropological fieldwork (although a lucky few of us do all too rarely get to go out and collect samples), molecular anthropology has already had, and is continuing to have, a major impact on our understanding of our evolutionary past—from the first demonstration of a surprisingly close relationship between humans and chimpanzees in the 1960s, to the mtDNA evidence for a recent African origin that developed in the 1980s, to the current fascination with whole genome sequences from Neandertals and other archaic humans.

Molecular anthropology can thus be considered a full-fledged, mature subfield of biological anthropology (alongside paleoanthropology, primatology, and demography), and therefore deserving of equal coverage in the curricula of university anthropology departments. However, the treatment of molecular anthropology in most undergraduate textbooks in biological anthropology or human evolution is often quite superficial and generally leaves a lot to be desired—while there are some good advanced books, there is nothing really comparable for the beginning student, who may have little in the way of any previous background in science. The present book is an attempt to remedy this situation by assuming no prior knowledge of genetics and by trying to focus on understanding the logic and reasoning behind various methods and findings, while omitting (or at least, placing less emphasis on) the technical details.

In addition to beginning students, it is hoped that this book will be useful to professionals from other fields (such as linguists or archaeologists) who want to know more about molecular anthropology and how it might inform their own work, as well as the interested layperson. The power of the molecular approach to anthropology lies in the fact that each of us carries within us a record of our past in the DNA that we have inherited from our ancestors, and the challenge is to learn how to read that record from the patterns of DNA variation in people today (supplemented, increasingly, by DNA extracted from fossils). Most people are intensely interested in human origins in general and their own origins in particular, and a whole industry now exists that will allow you to investigate your own genetic ancestry (for a suitable fee, of course!). But if you like, you can go beyond "personal genomics" and carry out your own investigations—while most of us will never have the opportunity to go on expeditions to dig up fossils or artifacts, you don't need your own laboratory to study genetic history. Anyone with a computer and a reasonably fast Internet connection can download genetic data from public repositories and freely available software to carry out various analyses (or, for the truly ambitious, write your own software), and voilà, you too can do molecular anthropology research. This book is thus also intended for anyone

interested in knowing more about what molecular anthropology is all about, as well as those who may be thinking about carrying out their own studies (but be forewarned that this is not a "how-to" book; you'll have to look elsewhere for step-by-step instructions—there are lots of resources on the Internet devoted to this sort of "armchair" molecular anthropology).

This book is loosely organized into three sections. The first six chapters are intended as introductory material for those who have never had any courses in genetics: Chapters 1 and 2 cover the basics of how genes are inherited, what they are, what they do, and how they do it; Chapter 3 introduces some basic properties about populations, including the important concept of effective population size; Chapter 4 sets up a simple (and highly unrealistic!) model of how genes behave in populations that nevertheless leads to some important insights; Chapter 5 makes the simple model of Chapter 4 more realistic by adding various evolutionary forces, with a focus on what happens to genetic variation within populations and genetic diversity between populations; and Chapter 6 covers some aspects of how genes themselves evolve.

The second section includes the next six chapters and provides an overview of the different types of genetic data and analyses that can be employed in molecular anthropology studies. Chapter 7 covers the various types of genetic markers that have been used and how they are analyzed in the laboratory, while Chapter 8 discusses issues that arise with sampling of populations (an important but often-overlooked aspect of molecular anthropology studies that can greatly impact the results) and Chapter 9 discusses the properties of different parts of the genome that are typically analyzed (which can also have a big impact on the results). The next three chapters focus on methods for analyzing genetic data, where the data come from populations (Chapter 10), which is the traditional approach, or from individuals (Chapter 11), which is a relatively new development made possible by new molecular methods; these two chapters focus largely on descriptive methods, while Chapter 12 is devoted to actually inferring demographic history from molecular data (i.e., estimating divergence times, changes in population size, etc.).

These first 12 chapters set the stage for the last eight chapters, which are devoted to what we have actually learned from molecular anthropology studies. We begin with what are (arguably) two of the most important contributions of the molecular approach to anthropology: namely, figuring out who is our closest living relative and just how close is the relationship (Chapter 13) and figuring out how our own species (modern humans) originated (Chapter 14). It turns out that the story of our origins in Chapter 14 is incomplete without the assistance of ancient DNA, and so Chapter 15 then discusses the various issues that arise with the analysis of DNA from fossils, and what we have learned. Hopefully, it is not giving too much away at this point to say that the genetic evidence strongly supports an origin of our species in Africa; Chapter 16 then discusses what we have learned from genetic evidence about the migration of modern humans from Africa, as well as two of the major subsequent migrations of modern humans: the colonization of the New World and the colonization of the Pacific.

Up to this point, the focus of the book is on demographic aspects of human history, that is, when did events take place, where did they take place, who did they involve, were there changes in population size, and so forth. But another very important aspect of our evolution is adaptation: what were the genetic changes that were selected for during our evolutionary past that allowed us to evolve to become modern humans, and what sorts of adaptations occurred subsequently as our ancestors spread across and out of Africa? Chapter 17 discusses species-wide selection, that is, selection for adaptations that are shared by all modern humans and thus can be thought of as those changes that made us human. In Chapter 18, we discuss local selection, that is, selection that occurred only in some populations due to their particular environment, climate, diet, diseases/parasites, and so forth; these can be thought of as adaptations that allowed us to successfully colonize more of the globe than any other species (with the exception of our parasites, of course!).

Chapter 19 turns to some aspects of genes and culture, in particular, the impact of cultural practices on patterns of genetic variation, as well as how we can use genetic analyses to make inferences about some cultural practices—one of the examples discussed in this chapter is a genetic approach to dating the origin of clothing (I kid you not!). The book ends with a final chapter that describes some of the other ongoing and likely future developments in molecular anthropology—a risky business, given the rapid rate of technological and computational advancements in this field. For example, nobody writing a textbook a few years ago would have predicted that in 2013 we would have high-quality, whole genome DNA sequences from Neandertals (and other archaic humans). It truly is an amazing time to be doing this sort of work, and I, for one, can't wait to see what we'll be able to do a few years from now.

In writing this book, I have in many places taken advantage of the fact that I have been actively involved in molecular anthropology research for more than 30 years and have been privileged to either participate in or have a ringside seat at some of the most significant advances in the field (e.g., the mtDNA and recent African origins research, and the analysis of DNA from archaic humans such as Neandertals). This can be

considered both a blessing and a curse. On the one hand, it is a blessing because I have drawn on my experiences and presented many results from my own studies, not because they are so much better than other studies but because by doing so, I can provide some behind-the-scenes insights into how such research is actually done and the decisions that have to be made along the way. Hopefully, the reader will thereby come away with a fuller appreciation of not only what we have learned from molecular anthropology about our origins and evolution but also how science in general is a process and not just an outcome.

On the other hand, drawing so heavily on my own research is a curse because of the potential biases that may creep in. While molecular anthropology is a science, in that we try to frame hypotheses that make predictions about genetic data that we can thereby test, it is a historical science, not an experimental science. We can't actually recreate the past—the only way we could ever know for sure what happened would be to invent a time machine and go back and directly observe the past—so we are left with making inferences about the processes and events that would most likely produce the patterns of genetic variation that we observe today. But this is inherently inexact—Occam's razor notwithstanding, the simplest explanation is not necessarily the true explanation—so there is plenty of room for different opinions and interpretations, which sometimes can get quite contentious! While I have tried to identify other points of view and make clear what is opinion versus what is "fact," it is nonetheless the case that not everybody would agree with everything in this book. Fortunately, it is a very simple matter to find alternative views by simply searching the Internet, so don't feel constrained by what is presented in this book.

There are a few people I'd like to thank for helping make this book a reality: Karen Chambers, my editor at Wiley, gets a special nod for guiding me all along the way and offering suggestions and encouragement (having a former student as your editor certainly is beneficial in the way your editor then treats you!); Stephanie Dollan for her able assistance; Rebecca Lim and Baljinder Kaur for handling the production; Rupak Kumar for handling the illustrations; and Sylvio Tüpke, Marike Schreiber, and Chloe Piot for their last-minute assistance with the illustrations. The ideas and interpretations expressed in this book are the product of interactions with many students and colleagues over the years, too numerous to mention. I have tried to give credit where credit is due, but I am sure I've overlooked or forgotten some of the details, and maybe even made a mistake or two, so corrections and constructive criticism are welcome and will be incorporated in future editions (should there be any). But the lion's share of the credit (and none of the blame) goes to Brigitte Pakendorf, who cajoled and persuaded me into thinking that maybe I actually could write a book. I leave it to the reader to decide if she was actually correct in her judgement.

CHAPTER 1

GENES: HOW THEY ARE INHERITED

Like begets like: dogs have puppies, cats have kittens, and humans have baby humans. Moreover, you tend to look more like your parents or other relatives than people you are not related to. The mechanics behind these simple statements—the laws of heredity—were first worked out by Gregor Mendel in the 1860s, who studied how variation in garden peas was transmitted from parents to offspring (Mendel 1865). But peas aren't so terribly interesting—and after all, this is an anthropology textbook—so we will use variation in humans to illustrate the mechanics of inheritance. The variation we will use is the ABO blood group system, but before explaining how the ABO blood groups are inherited, you first need to know something about blood.

BLOOD AND ABO BLOOD GROUPS

Suppose you stick a needle with a syringe into a vein, withdraw a few ccs (cubic centimeters—a cc is about 20 drops or so) of blood, squirt the blood into a test tube, and let it sit. After 30 minutes or so, the blood will have spontaneously formed a clot—all it takes is exposure of the blood to air to initiate clotting. Remove the clot and what is left behind is a clear, yellowish fluid called **serum**. If you instead add a chemical to the test tube that inhibits clotting and spin the blood at high speed in a centrifuge, you will find that the blood has separated into different components (Figure 1.1). At the bottom are the red blood cells (RBCs, also known as **erythrocytes**), which transport oxygen around the body. Immediately on top of the RBCs is a ghostly white layer, sometimes referred to as the buffy coat, that consists of white blood cells (also known as **lymphocytes**), which are important for protecting the body from invading cells. And on top of the white blood cells is a clear, yellowish fluid called **plasma**. Plasma is like serum, except plasma also contains the various factors that are involved in blood clot formation.

Suppose now we take serum from one person and mix it with RBCs from another person and do this for many different people. Sometimes nothing will happen, but sometimes the RBCs will clump together (**agglutinate**). Agglutination is entirely different from clotting (Figure 1.2). You may think that mixing blood components from different people is a strange thing to do, but in fact Karl Landsteiner won a Nobel Prize for doing just that. During the nineteenth century, physicians began giving blood transfusions to people who had lost life-threatening quantities of blood through injury or illness. Seems reasonable enough—someone needs more blood, so give them blood from somebody else—and indeed, sometimes the blood transfusion recipients recovered spectacularly. But sometimes they actually got much sicker from the transfusion, to the point of even dying, and nobody knew why this would happen. Landsteiner, an Austrian physician, took it upon himself to figure out why such adverse reactions to blood transfusions occurred. Through his mixing experiments, he discovered that people's blood could be classified into four groups (Landsteiner 1900), corresponding to what are now known as blood groups A, B, AB, and O. Mix together blood from people with the same blood group and nothing happens. But mix together blood from a group A person with blood from a group B person and you get agglutination—and if you do this in a blood transfusion, clumps of agglutinated cells will form in the veins, blocking small capillaries and leading to tissue death, which is bad news indeed.

So what causes agglutination? It turns out that RBCs carry on their surface substances called **antigens**, and

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

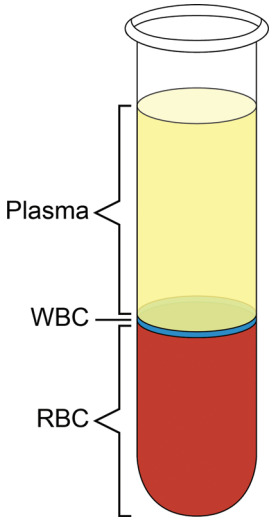


FIGURE 1.1

The components of blood, after adding an anticoagulant, followed by centrifugation. RBC, red blood cells; WBC, white blood cells.

these antigens cause the formation of substances in the serum called **antibodies**, which bind to antigens. Each antibody has two binding sites for its particular antigen, and there are many copies of each antigen on each RBC. So, mix together RBCs with serum containing antibodies against an antigen on those RBCs, and you get lots of antibodies binding to lots of RBCs, resulting in agglutination. But if the serum does not contain antibodies against the antigens on the RBCs, then there is no agglutination.

Table 1.1 lists the antigens present on the RBCs and the antibodies present in the serum of the A, B, AB, and O blood groups (for those of you who have seen blood groups with + or −, such as A+ or B−, don't worry, we'll get to that later in the chapter). The O blood group can be thought of as a "null" blood group, in that there are no O antigens or anti-O antibodies. Note that if you have a particular antigen on your RBCs, you don't have antibodies against that antigen—otherwise you would be agglutinating your own blood cells, which would be very bad news indeed (however, there are diseases known in which the body starts making antibodies against its own antigens; such diseases are known as **autoimmune diseases** and examples include lupus and some types of arthritis). Note that people with blood type O are known as "universal donors," because their RBCs lack A or B antigens and hence can be safely transfused into people of any blood type—that's why you often hear emergency room physicians on TV shows shouting for type O blood when a patient comes in who needs blood immediately. Conversely, people of blood type AB are known as "universal recipients," because they can receive RBCs of any blood type in a transfusion, as they lack anti-A and anti-B antibodies.

TABLE 1.1 ■ Antigens and antibodies for the ABO blood groups

Blood group	Antigens on RBCs	Antibodies
A	A	anti-B
B	B	anti-A
AB	A,B	none
O	None	anti-A, anti-B

RBCs, red blood cells.

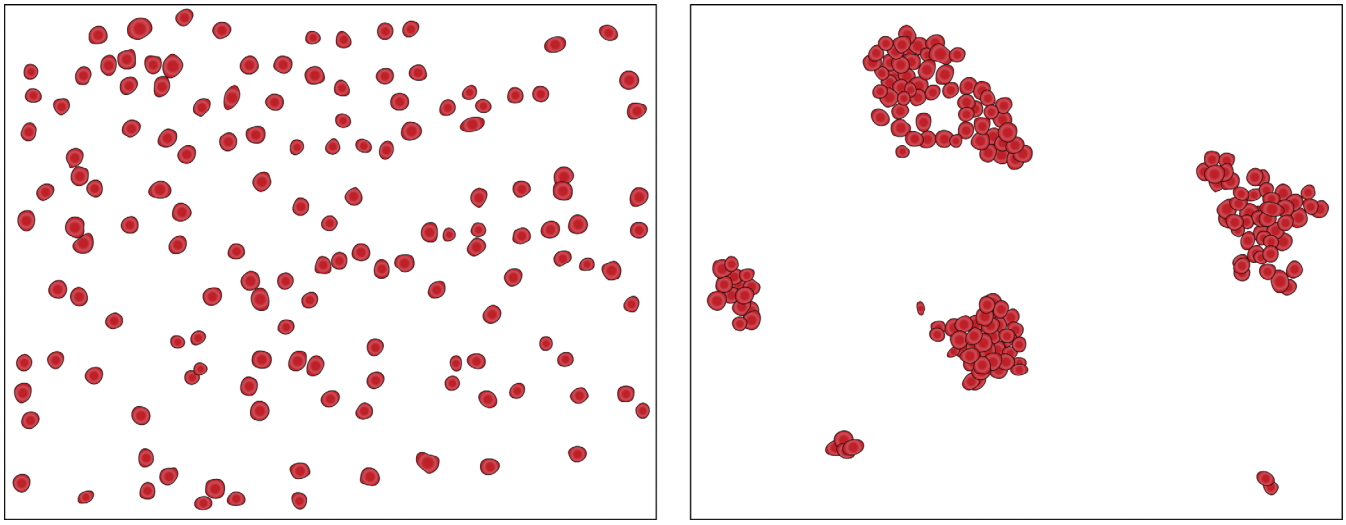


FIGURE 1.2

Left, a version of red blood cells that have not agglutinated. Right, a version of red blood cells that have agglutinated.

INHERITANCE OF ABO BLOOD GROUPS

Now that you know something about ABO blood groups, we can go into how they are inherited. First, some facts and terminology. Humans are **diploid**, meaning that each **gene** is present in two copies (for now, just think of a gene as the instructions for doing something, as in "the gene for the ABO blood groups"; in the next chapter, we'll see what genes actually are). One copy is inherited from the mother, through the egg, and one copy is inherited from the father, through the sperm. Any particular gene can come in different forms, or variants, and these are called **alleles**. For the ABO blood group gene, there are three alleles, namely, the A allele, the B allele, and the O allele. And since everyone has two alleles, there are six possible combinations of alleles; the pair of alleles that you have is your **genotype**. For three genotypes, the two alleles are the same (namely, AA, BB, and OO), and these are called **homozygous genotypes** or **homozygotes**. For the other three genotypes, the two alleles are different (namely, AB, AO, and BO), and these are called **heterozygous genotypes** or **heterozygotes**. The astute reader may wonder how it is that six different genotypes result in just four different blood groups. The actual blood group, or **phenotype**, associated with each genotype is shown in Table 1.2. Note that both the AA genotype and the AO genotype result in blood type A, and both the BB genotype and the BO genotype result in blood type B, thereby explaining how six different genotypes result in just four different blood groups.

The ABO blood groups also nicely illustrate the concept of **dominant** versus **recessive** alleles. If the heterozygote for two alleles exhibits exactly the same phenotype as the homozygote for one of the alleles, then that allele is said to be dominant, and the allele that does not exhibit a phenotype in the heterozygote is said to be recessive. Thus, since the AO genotype results in exactly the same phenotype (blood group) as the AA genotype, the A allele is dominant with respect to the O allele, and the O allele is recessive with respect to the A allele. Similarly, the B allele is dominant with respect to the O allele, and the O allele is recessive with respect to the B allele, because the phenotype of the BO heterozygote is exactly the same as that of the BB homozygote. What about the A and B alleles—which is dominant and which is recessive with respect to each other? To figure this out, look at the phenotype (blood group) associated with AB heterozygotes. It turns out that AB heterozygotes have a different phenotype than either AA or BB homozygotes—they are type AB. We therefore say that the A and B alleles are **codominant** with respect to each other (other terms you may come across, such as **partial dominance** or **incomplete dominance,** mean basically the same thing as codominance: the heterozygote has a different phenotype than either homozygote).

TABLE 1.2 ■ ABO blood group genotypes and corresponding phenotypes

Genotype	Phenotype (blood type)
AA	A
AO	A
BB	B
BO	B
AB	AB
OO	O

Note that the dominance relationship is a property of a pair of alleles, not of a single allele, and, therefore, can vary depending on which pair of alleles are considered. For example, it would be incorrect to simply say that the A allele is dominant, because even though it is dominant with respect to the O allele, it is codominant with respect to the B allele. Determining the dominance relationships of a pair of alleles simply involves comparing the phenotype of the heterozygote to the phenotype of each homozygote. If the heterozygous phenotype matches one of the homozygotes, then that allele is dominant and the other is recessive. If the heterozygous phenotype differs from both homozygotes, then the alleles are codominant.

A lot of terminology was introduced in the previous paragraphs—but if you want to walk the walk, you've got to be able to talk the talk. So, the sooner you become conversant with the terminology—at the very least, know what is meant by gene versus allele, genotype versus phenotype, homozygote versus heterozygote, and dominant versus recessive versus codominant—the better. Now, how are ABO blood groups transmitted from parents to offspring? Recall that humans are diploid, with two ABO blood group alleles, one inherited from the mother and one inherited from the father. This means that the mother's egg and the father's sperm are **haploid**, carrying one allele each instead of the usual two alleles. If the parent is homozygous, then all of the **gametes** (eggs for women, sperm for men) produced by that parent will carry the same allele. But if the parent is heterozygous, then on average half of the gametes will carry one allele, and half will carry the other allele. Knowing the genotypes of the mother and the father, we can then predict the genotypes of the offspring. For example, suppose one parent has the AA genotype and the other parent has the AB genotype. The AA parent will produce only A gametes, while the AB parent will produce 50% A gametes and 50% B gametes. Thus, we expect that any child of these parents has a 50% chance of being genotype AA and a 50% chance of being genotype AB. Moreover, if we look at lots and

		Father	
		A	O
Mother	A	AA	AO
	O	AO	OO

FIGURE 1.3

Punnett square illustrating the ABO blood group genotypes expected among the children when both parents have the AO genotype.

lots of children where one parent is AA and the other is AB, we expect about half the children to have genotype AA and half to have genotype AB.

In this example, the children end up having the same genotypes and blood groups as the parents. However, this need not always be the case. A convenient way of diagramming the expected outcome of any type of mating is the **Punnett square**, imaginatively named after its inventor, the geneticist Reginald Punnett. An example of a Punnett square is shown in Figure 1.3 for the case when both parents are of genotype AO (hence blood type A). In this situation, 25% of the children are expected to be genotype OO, and hence blood type O. So, having a child of blood type O when the parents are both type A (or both type B, or one is type A and one is type B) need not be a cause for concern on the part of the father, as genetics shows how this can arise. However, genetics cannot so easily explain a child of blood type A or B when both parents are blood type O (do the Punnett square if this is not immediately obvious to you), so in such cases, the mother would have some explaining to do to the father!

The idea that gametes carry only one allele, and that a heterozygous parent produces gametes carrying either allele in equal frequency, is the basis of Mendel's First Law of Segregation (i.e., alleles **segregate** into gametes). There are two important consequences. First, offspring are produced by the random union of gametes, hence the outcome of one mating has no influence on the outcome of subsequent matings. Suppose a genotype AA parent and a genotype AB parent have an AA child. The chance that the next child is genotype AB is still 50%. Suppose these same parents have 10 children, all of genotype AA. We may now wonder if perhaps we haven't made a mistake in our genotyping of the parents, but assuming the genotypes are correct, then the chance that the eleventh child is genotype AB is still just 50%. There is no "memory" to the system, no compensating for prior events—predicting the genotype of a child is subject to the same laws of chance as flipping a coin.

The second important consequence of Mendel's First Law of Segregation is that inheritance is **particulate**. That is, whatever genes are (and remember, all the mechanics of how genes are inherited were worked out long before anybody knew what genes actually are), they behave as discrete particles. Prior to the rediscovery of Mendel's work, it was generally assumed that inheritance was **blending**: genes were thought to behave like blood (thus, all the emphasis on people's bloodlines), so the characteristics of the genes in the parents would become mixed in the children. And the children would in turn transmit these mixed characteristics to their children, and so forth.

Blending inheritance may sound reasonable, but it posed a big problem for Darwin's theory of evolution. Darwin proposed that individuals with characteristics that enhanced their survival or fertility would transmit those characteristics to their offspring, thereby increasing the frequency of such advantageous characteristics in subsequent generations. But if in each generation the advantageous characteristics are blending with the less-advantageous characteristics, then it is hard to see how advantageous characteristics can increase in frequency. It's like mixing paint—mix red and white paint together and you will get pink paint, and no matter how much more red or white paint you add, you still end up with various shades of pink. Indeed, Darwin spent a long time grappling with this issue and never came up with a satisfactory answer.

However, the idea that genes behave as particles neatly solves the problem. Suppose an individual of ABO blood group genotype AA (hence, blood type A) has a child with an individual of genotype OO (hence, blood type O). The child (genotype AO, blood type A) grows up and then marries an AA individual (blood type A) and has one child who is genotype AO (blood type A). Imagine that this continues for 10 generations, with each generation producing an AO individual who marries an AA individual and has an AO child. Now, after 10 generations of only blood type A in this family, suppose in the eleventh generation the AO individual marries an individual with genotype OO (blood type O) and they have a child with genotype OO. This child will have the O blood type—the fact that the O allele came from a long line of individuals of genotype AO, who were all blood type A, does not change what that O allele does when it is now paired with another O allele. It's as if we mixed red with white paint to get pink paint, but then we can get pure red or pure white paint back out of the mixture.

INHERITANCE OF MORE THAN ONE GENE: ABO AND RHESUS BLOOD GROUPS

To illustrate the mechanics of inheritance for more than one gene, we will use the second blood group

to be discovered, so first you need to know something about this blood group. Although blood transfusion success increased markedly with the recognition of the importance of the ABO blood groups, serious reactions after a blood transfusion still happened, even when the donor and the recipient were matched for ABO blood type. Moreover, it became apparent that a disease called **hemolytic disease of the newborn** (HDN) was due to antibodies from the mother crossing the placenta and attacking an antigen on fetal RBCs. Hemolytic disease of the newborn is quite serious as it can result in severe anemia, jaundice, and even death of the newborn—and again, HDN was observed even when there was no ABO blood group incompatibility between mother and child. These observations lead to the discovery of the second human blood group, namely, the **rhesus** (Rh) blood group—so named because it was initially thought that the factor causing blood transfusion reactions and HDN was identical to an antigen identified first on rhesus monkey RBCs and then shown to also occur on human RBCs (Landsteiner and Wiener 1940). Actually, we now know that the HDN-causing factor and the antigen on rhesus monkey RBCs are not the same, but the name stuck.

The rhesus blood group is a very complex system but can be simplified into two major alleles, Rh+ and Rh−. The Rh+ allele is dominant to the Rh− allele, so there are two blood types (phenotypes): Rh positive (corresponding to genotypes Rh+/Rh+ and Rh+/Rh−) and Rh negative (corresponding to genotype Rh−/Rh−). These are the source of the + and − that is added on to the ABO blood type, for example, A+ means that person is ABO blood type A and Rh blood type positive, while O− means that the person is ABO blood type O and Rh blood type negative.

People who are Rh positive have Rh+ antigens on their RBCs but no Rh antibodies; people who are Rh negative do not have Rh antigens on their RBCs and hence can make anti-Rh+ antibodies if exposed to Rh+ RBCs. Note that this is the usual way that antibodies work: you only make the antibodies after you are exposed to the antigen. If you are Rh negative, you won't make anti-Rh+ antibodies until you are exposed to RBCs with the Rh+ antigen. So, an Rh− person could be transfused with Rh+ blood without suffering any ill effects—by the time any anti-Rh+ antibodies are made, the transfused Rh+ RBCs will no longer be present. A second such transfusion of Rh+ blood, however, would be bad news, because now anti-Rh+ antibodies will already be present from the first transfusion and they can agglutinate the transfused Rh+ RBCs. Note also that the ABO antibodies are an apparent exception to the rule that you make antibodies only after you are exposed to antigens, since you are born with antibodies to the ABO antigens that you do not possess. What seems to happen is that chemical

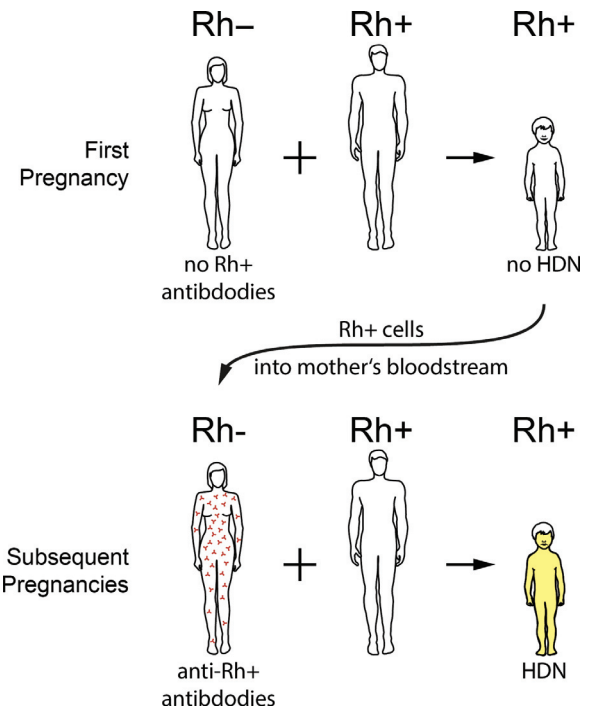


FIGURE 1.4

The circumstances leading to HDN. See text for details. HDN, hemolytic disease of the newborn.

substances that are similar to the ABO antigens are so widespread in nature (they are simple sugars that are commonly found in the environment) that exposure occurs somehow in the womb, resulting in production of the antibodies even before birth.

So, how does HDN arise? Hemolytic disease of the newborn occurs under the following circumstances (Figure 1.4): when an Rh− mother has an Rh+ child (which can happen when the father is Rh+), ordinarily nothing happens to the first such child. However, fetal cells typically do cross the placenta and get into the mother's bloodstream. If the mother is Rh+, nothing will happen, as she will not develop anti-Rh+ antibodies, but an Rh− mother will react against the Rh+ antigens on the fetal RBCs and develop anti-Rh+ antibodies. If the Rh− mother then subsequently becomes pregnant with another Rh+ child, the mother's anti-Rh+ antibodies can cross the placenta and attack the fetal RBCs that carry the Rh+ antigens, resulting in HDN. Untreated HDN results in death in about one-third of the cases, so this is a serious matter; affected infants usually need blood transfusions and treatment for jaundice (caused by excess levels of hemoglobin due to the destruction of fetal blood cells) immediately.

Fortunately, there is a simple and effective means of preventing HDN, and that is to give the mother an injection of concentrated anti-Rh+ antibodies shortly after the birth of the first child (and after any

subsequent children). These antibodies coat any Rh+ fetal RBCs that make it into the mother's bloodstream, thereby preventing the mother's immune system from making her own Rh+ antibodies. This injection usually goes by the name "Rhogam," so those of you who have experienced pregnancy either directly or via a pregnant partner and wondered about this Rhogam injection, now you know.

Incidentally, there are more than 30 different blood group systems known. However, the ABO and Rh blood groups are by far the most important because of their role in blood transfusions and HDN. That's why most of you probably know your ABO/Rh blood type but not your Lewis, Kell, or any other blood type. Still, these other blood groups sometimes pop up in cases involving adverse reactions to blood transfusions or HDN. In such cases, the first course of action is to check the ABO/Rh blood type, and if these cannot explain what is going on (e.g., a case of HDN where the mother is Rh+), then some other blood group must be involved, and in fact this is how most of these other blood groups were discovered.

The inheritance of the Rh blood type alone is quite simple, as Rh+ is dominant to Rh−. But what about the inheritance of both ABO and Rh blood type? Consider the following example, shown in Figure 1.5, where one parent is type AB− and the other parent is type O+, and we want to know what to expect for the children. The first step is to figure out the genotypes of the parents. The AB− parent can have only the A/B, Rh−/Rh− genotype, but the O+ parent can have one of two possible genotypes: O/O, Rh+/Rh+ or O/O, Rh+/Rh−. Without any further information, we don't know which genotype this person has, but suppose we know that one of this person's parents was O+ and the other was O−. Then we know that this person must have inherited an Rh− allele from the O− parent; hence, the genotype must be O/O, Rh+/Rh−. As shown in Figure 1.5, we then expect four blood types among the children: A+, A−, B+, and B−. And, we expect these to occur in equal frequency, so there is a 25% chance of any one child having any one of these blood types.

Father

Mother	O+	O−
A−	AO, +/−	AO, −/−
B−	BO, +/−	BO, −/−

FIGURE 1.5

Punnett square illustrating the ABO and Rh blood group genotypes expected among the children of a mother with the AB, Rh−/Rh− genotype and a father with the OO, Rh+/Rh− genotype. The genotypes at the ABO and Rh genes assort independently.

We have just demonstrated **Mendel's Second Law of Independent Assortment**: alleles from different genes assort independently into gametes. That is, if you go back to the example in Figure 1.5, you see that for just the ABO gene, from Mendel's First Law, there is a 50% chance of a child with blood type A and a 50% chance of a child with blood type B. And, by the same reasoning, if you consider only the Rh gene, there is a 50% chance of an Rh+ child and a 50% chance of an Rh− child. To get the probability for both the ABO and Rh blood types, multiply the separate probabilities: the chance of an A+ child, for example, is 50% of 50%, or 25%. Independent genes behave independently, so the probability of having a child of a particular genotype for two (or more) genes is obtained by multiplying the probabilities for each genotype—just as you would do if you wanted to know the probability of getting both a head by flipping a coin and a six on a roll of a die (which would be 1/2 times 1/6, or 1/12).

Let's look at another example of Mendel's Second Law, this time using some (slightly modified) actual data. Table 1.3 shows some data from families with **elliptocytosis**, a hereditary blood disorder in which a large fraction of the RBCs have an elliptical shape rather than the usual disc shape. In severe cases, the afflicted individuals suffer from anemia, as the abnormal RBCs break down prematurely. Elliptocytosis is a partially dominant disease, meaning that heterozygotes show some of the symptoms, while homozygotes are even more strongly afflicted. Table 1.3 also includes the Rh blood type information, and for reasons that will become clear in just a minute, the data in Table 1.3 are specifically chosen from families where one parent is heterozygous for both Rh and for elliptocytosis (i.e., Rh+/Rh−, Ep+/Ep−, using Ep+ to designate the disease-associated allele and Ep− to designate the

TABLE 1.3 ■ Observed number of offspring who are Rh+ or Rh− and either afflicted with elliptocytosis (Ep+) or not (Ep−) in families as discussed in the text[a]

Phenotype	Observed number
Rh+, Ep+	34
Rh+, Ep−	3
Rh−, Ep+	4
Rh−, Ep−	32

[a]Data are taken from Lawler, S.D., and Sandler, M., *Annals of Eugenics* 18:328–334 (1954); as the data come from a variety of families with a variety of genotypes, I have taken the liberty of tabulating the data as if they all came from families with the same parental genotypes, in order to make things simple. The key observations (66 offspring of the "major" or parental types and seven offspring of the "minor," or recombinant types) are as reported by Lawler and Sandler and led to the conclusion of linkage between the rhesus blood group and elliptocytosis loci.

"normal" allele) while the other parent is homozygous for the recessive alleles at both genes (i.e., Rh−/Rh−, Ep−/Ep−). Note that in such families, according to Mendel's Second Law, we expect four possible genotype combinations that should occur in equal frequencies, with the associated phenotypes as follows:

25% Rh+/Rh−, Ep+/Ep−, which are Rh positive and affected with elliptocytosis

25% Rh+/Rh−, Ep−/Ep−, which are Rh positive and not affected with elliptocytosis

25% Rh−/Rh−, Ep+/Ep−, which are Rh negative and affected with elliptocytosis

25% Rh−/Rh−, Ep−/Ep−, which are Rh negative and not affected with elliptocytosis

(Do the Punnett square if this isn't obvious to you). And the results? As you can see in Table 1.3, the observed results are quite different from those expected by Mendel's Second Law of Independent Assortment.

So, what is going on here? One possibility is that nothing of any significance is going on, and what we have observed is simply a chance deviation from the expected frequencies. After all, we don't expect to get exactly 25% of each phenotype, just as if we flip a coin 10 times, we don't expect to get exactly five heads and five tails. But how likely are we to get the results in Table 1.3, if we actually expect 25% of each combination? This is a question for statistics, and rather than run the risk of scaring off readers now, we'll put off the discussion of statistical tests to Chapter 4. For now, just take it on faith that it is extremely unlikely that we would obtain the data in Table 1.3 if the true frequencies really were 25% of each phenotype.

If the data don't fit our expectations, then either there is something wrong with the data or there is something wrong with our expectations. In this case, the problem is with the expectations, because it turns out that the elliptocytosis and rhesus blood group genes are an example of a very important and well-known exception to Mendel's Second Law of Independent Assortment. This exception involves genes that are located close to one another on the same **chromosome**. We'll learn more about chromosomes in the next chapter; for now, all you need to know is that chromosomes are the physical structures within cells that contain genes. Chromosomes come in pairs, with one member of each pair inherited from the mother and the other inherited from the father. Humans have 23 pairs of chromosomes in each cell (Figure 1.6). So, genes have specific, physical locations on chromosomes, which is where the term **locus** comes from, as a synonym for a gene—we can talk about the ABO blood group gene, or the ABO blood group locus. And

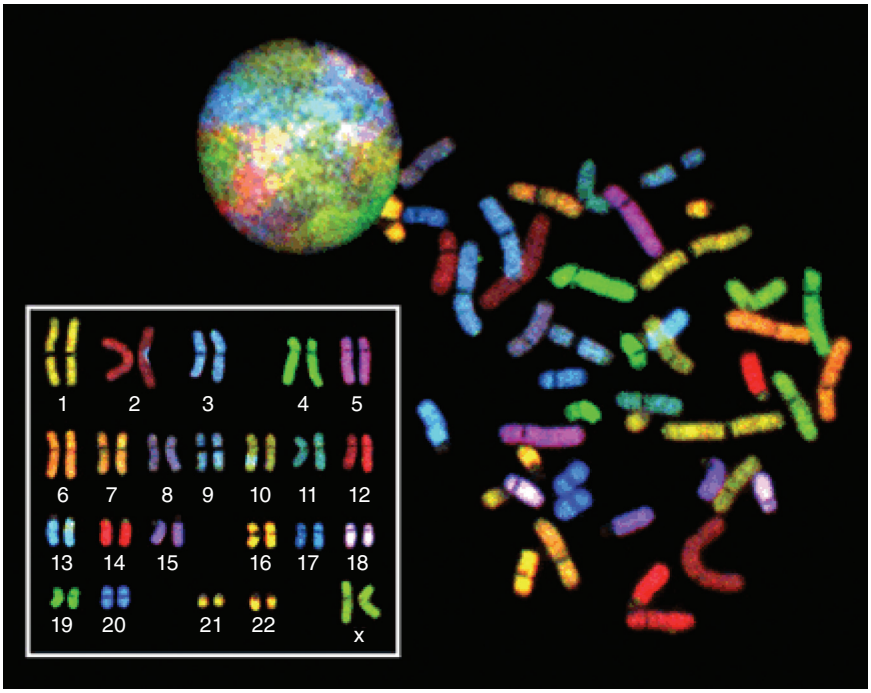


FIGURE 1.6

Example of a human karyotype, showing the 23 pairs of chromosomes. In this example, from a female, each chromosome has been stained with a different fluorescent dye; this is known as a spectral karyotype. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Sky_spectral_karyotype.png).

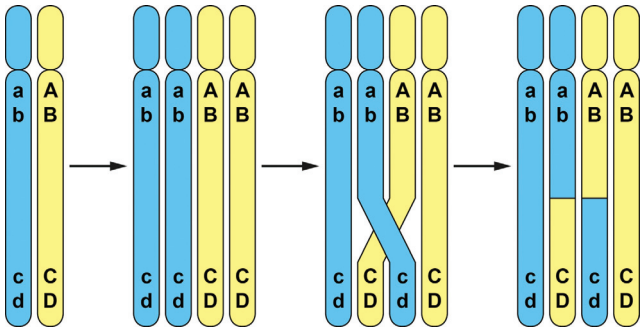


FIGURE 1.7

Recombination between chromosomes during meiosis, resulting in the exchange of chromosome segments. In this example consisting of 4 genes, each with two alleles, the individual inherited one chromosome with alleles abcd from one parent, and another chromosome with alleles ABCD from the other parent. These chromosomes duplicate and then two of them undergo recombination, with the result that there are two parental chromosomes (abcd and ABCD) and two nonparental or recombinant chromosomes (abCD and ABcd).

the key point is that genes located near one another on the same chromosome are **linked**, and the alleles that are on the same chromosome will be inherited together more often than predicted by chance. You might think that alleles on the same chromosome will always be inherited together, but such is not the case: during **meiosis** (the process of forming gametes—eggs and sperm), there is exchange (**recombination**) of segments between the two copies of each chromosome (Figure 1.7). In other words, you have two copies of each of your chromosomes, one you inherited from your mother and one you inherited from your father. But as shown in Figure 1.7, when you have children, the set of haploid chromosomes that you transmit to them will not be intact copies of either your maternal or paternal chromosomes. Instead, each chromosome you transmit to your children will contain some segments from your paternal copy and some from your maternal copy of that chromosome. However, each chromosome that you transmit will be a faithful copy in that all genes will be present and in the correct order (barring the rare chromosomal change that results in duplications of segments, loss of segments, or a different order of segments—these sorts of events will be discussed later). The results in Table 1.3 (which depart from the expected 25% of each combination of Rh and elliptocytosis alleles) are most simply explained if the genes for elliptocytosis and the Rh blood group are linked (located close together on the same chromosome)—which indeed they are.

The concept of **linkage** is extremely important, as we'll see in a minute, but first some technical points about detecting linkage. Note that in Table 1.3 we focused on families where one parent was known to be heterozygous for both Rh and Ep, and that is a general requirement: in order to detect whether two loci are linked or not, at least one parent must be heterozygous for both loci (i.e., doubly heterozygous). This is so we can distinguish between **parental** and **nonparental** (or more accurately, **recombinant**) gametes produced by the heterozygous parent, as shown in Figure 1.7. When an individual is homozygous for one (or more) of the loci in question, then parental and recombinant types cannot be distinguished from one another (if this is not obvious, make the genotypes in Figure 1.7 homozygous instead of heterozygous and see whether you can distinguish parental from recombinant gametes). The parent who is doubly heterozygous is said to be the **informative** parent, because then we can tell whether or not recombination has occurred in the gametes produced by this parent. Note that in principle, there are two possible associations between the alleles at the two loci in the informative parent (assuming that the loci are indeed linked): in the case of the Rh and Ep loci, the informative parent could have the Rh+ and Ep+ alleles on one chromosome and the Rh− and Ep− alleles on the other chromosome, or the informative parent could have the Rh+ and Ep− alleles on one chromosome and the Rh− and Ep+ alleles on the other chromosome. The particular combination of associated alleles is known as the **phase**; note that the phase can be different in different individuals, so you have to be careful when combining data from different families. The phase can sometimes be determined if you have data from the parents of the informative parent. For example, if the father of the informative parent is Rh−/Rh−, Ep+/Ep−, and the mother is Rh+/Rh−, Ep−/Ep−, then the informative parent has one Rh−/Ep+ chromosome and one Rh+/Ep− chromosome. If this isn't immediately obvious, note that the Ep+ allele had to come from the father, who is Rh−/Rh−, and so the father contributed an Rh−/Ep+ chromosome. Similarly, the Rh+ allele had to come from the mother, and so the mother contributed an Rh+/Ep− chromosome. Otherwise, you can compute how likely you are to observe the number of offspring of each parental/recombinant type, assuming each of the possible phases for the informative parent—but I ask you to take this on faith, as the details of this sort of computation are beyond the scope of this book. Determining the phase has other applications when it comes to making inferences about the demographic history of populations, and we will return to this topic in Chapter 9.

Recombination is a remarkable process that generates new genetic variation, in terms of shuffling

around maternal and paternal segments of chromosomes to create new combinations of alleles. Moreover, the amount of recombination is roughly proportional to the physical distance between linked genes: the alleles for genes that are located very close together on the same chromosome will tend to be inherited together, while for genes that are far apart on the same chromosome, there is so much recombination that their alleles will be inherited independently, as if they were on different chromosomes. Hopefully, this makes sense: the greater the distance between two genes, the more opportunity there is for one (or more) recombination events to occur between them. Think of it as placing two marks on a piece of string, then cutting the string at some random location. If the two marks are close together, only rarely will you cut the string in between them, but if the two marks are near the opposite ends of the string, then you'll almost always cut between them.

Linkage is of particular importance because it allows disease genes to be **mapped** (located on a chromosome) and ultimately identified by looking in families for the cosegregation of **marker genes** (genes whose chromosomal location are known) with the disease. Even just knowing about the linkage relationships of a disease gene can provide some useful information. For example, in the case of elliptocytosis, extensive family studies showed that some cases of elliptocytosis showed linkage to the Rh blood group locus (as in the example in Table 1.3) but others did not (Morton 1956). So, there must be more than one gene which, when mutated, can cause elliptocytosis—linkage studies thus provided some of the first evidence that what appears to be the same genetic disease can have different underlying causes. Linkage is also an important concept behind some strategies for identifying genes that have been subject to recent positive selection, as will be discussed in more detail in Chapter 18.

SEX CHROMOSOMES

There is an important extension to Mendel's First Law, which applies to genes found on the **sex chromosomes**. The members of each pair of chromosomes are physically indistinguishable for 22 of the 23 pairs of chromosomes in humans (Figure 1.6), and these are the **autosomes**, numbered from 1 to 22. The remaining pair are the sex chromosomes, dubbed X and Y, which are quite different; females have two X chromosomes while males have one X chromosome and one Y chromosome. Females thus produce gametes (eggs) carrying an X chromosome, while for males, 50% of the gametes (sperm) carry an X chromosome and 50% carry a Y chromosome. This accounts for the expected 50:50 male:female sex ratio and moreover makes clear that the responsibility for the determination of the sex of a child lies with the father, not with the mother—somebody should have informed Henry VIII before he lopped off the heads of various wives for failing to deliver a son!

The X and Y chromosomes differ greatly in size (Figure 1.6) and gene content; the X chromosome is much larger than the Y chromosome and has on the order of a thousand genes, while the Y chromosome has only about a dozen genes, mostly involved in male fertility. Importantly, the genes on the X chromosome thus do not have a corresponding copy on the Y chromosome, so males, with just one X chromosome, are said to be **hemizygous** for genes on the X chromosome. This means that the phenotype associated with a recessive allele at an **X-linked gene** (on the X chromosome) will always be manifested in males with that allele. For example, there are X-linked, recessive alleles that cause red–green colorblindness. A female who is heterozygous, having one normal color vision allele and one color blindness allele, will herself have normal color vision, because color blindness is recessive. However, if she has children with a male with normal color vision, there is a 50% chance that a son will be color-blind, but none of the daughters will be color-blind (as shown in Figure 1.8). Such X-linked recessive traits will, therefore, occur more often in males than in females. In fact, in order for a female to manifest a sex-linked recessive trait, she must inherit an X chromosome from her father who carries the recessive allele (do the Punnett square if this isn't obvious), so her father must also manifest the trait. Some X-linked traits are so debilitating that males with the trait hardly ever reproduce, and so these traits tend to occur only in males. Examples of such traits include hemophilia (which, until recently, was invariably lethal before affected men reached reproductive age) and some forms of mental retardation.

		Father	
		X^N	Y
Mother	X^N	X^NX^N	X^NY
	X^c	X^NX^c	X^cY

FIGURE 1.8

Punnett square illustrating the genotypes expected in the children where the mother is heterozygous for the colorblindness gene (X^N/X^c, where X^N is the allele for normal color vision and X^c is the colorblindness allele) and the father has normal color vision (X^N/Y). In such families, half of the male children are expected to have color blindness.

DETERMINING HOW TRAITS ARE INHERITED: PEDIGREE ANALYSIS

Given a particular trait of interest, how do we figure out how it is inherited? If we were interested in garden peas (or fruit flies or mice or other commonly used experimental organisms), then it would be simple: select individuals who differ in the trait, have them mate, and see what happens in the offspring and subsequent generations. With humans it's more complicated: it isn't ethical (or practical) to select people and have them mate, so we have to rely on what nature provides, namely, we analyze families where at least one individual has the trait of interest. This type of analysis is called **pedigree analysis**. Consider the example in Figure 1.9, which is a diagram of three generations of a family. To figure out how the trait is inherited, focus on the following questions: (1) do people with the trait have at least one parent with the trait; and (2) are there equal numbers of males and females with the trait? In the example in Figure 1.9, all people with the trait have parents with the trait, and there are roughly equal numbers of males and females with the trait. These are the hallmarks of **autosomal dominant** inheritance, where autosomal means that the gene is on one of the 22 pairs of physically identical chromosomes (autosomes) and dominant means that people with the trait can be either heterozygous or homozygous for the responsible allele (as, e.g., people with ABO blood type A can have either the AO or the AA genotype). Hopefully, it is clear by now why people with an autosomal dominant trait have a parent with the trait: if you have the autosomal dominant trait, you have at least one allele for the trait, which you must have inherited from one of your parents, who then must also have the trait. Armed with this knowledge, we can assign genotypes to the individuals as shown in Figure 1.9: if we designate the (dominant) allele for the trait **A** and the (recessive) allele for the absence of the trait **a**, then all of the individuals with solid symbols have the **Aa** genotype and everyone else has the **aa** genotype. Moreover, we can predict that if individual X in the figure (who has the trait) has a child, then there is a 50% chance that the child will have the trait. And, if individual Y (who lacks the trait) has a child with someone who also doesn't have the trait, then there is a 0% chance that their child will have the trait—even though individual Y has two sisters, a father, and aunt, and a grandfather with the trait (if either of these statements isn't immediately obvious, do the Punnett square!).

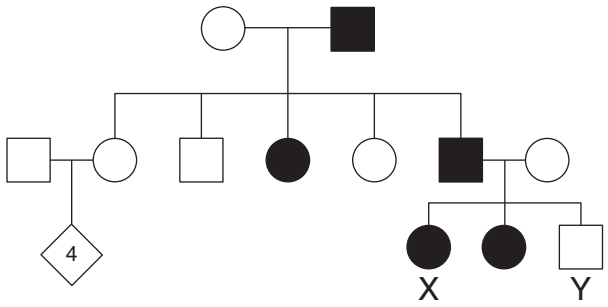


FIGURE 1.9

Pedigree illustrating autosomal dominant inheritance. Squares are males, circles are females, horizontal lines between a square and a circle indicate matings, and vertical lines indicate offspring. Solid symbols indicate individuals with the trait. The diamond with a 4 indicates four children of unknown sex without the trait.

However, there are important exceptions to these general statements about autosomal dominant inheritance. For example, achondroplastic dwarfism (a type of dwarfism characterized by a long, narrow trunk and short arms and legs) is an autosomal dominant trait in humans, and yet about 80% of achondroplastic dwarfs are born to parents of normal stature. But a hallmark of autosomal dominant inheritance is that people with the trait have a parent with the trait, so how can this be? It turns out that most cases of achondroplastic dwarfism are due to new **mutations**, not to inheritance of the allele for dwarfism from a dwarf parent. In Chapter 2, we will discuss how mutations occur. For now, just realize that most mutations are very rare, but for traits that are extremely harmful or otherwise greatly reduce a person's chances of having children, most cases of children with such traits do indeed reflect new mutations (for achondroplastic dwarfs, there is reduced fertility and often complications with pregnancy, which tends to limit the number of children they have).

Now let's consider the pedigree in Figure 1.10 and ask the same questions: do people with the trait have a parent with the trait; and are there roughly equal numbers of males and females with the trait? Here we see that both males and females have the trait, but people with the trait do not have a parent with the trait. These are the characteristics of **autosomal recessive** inheritance. The idea is that in order to exhibit an autosomal recessive trait, by definition a person must be homozygous for the relevant allele. And the most likely way for that to happen is for two heterozygotes to have a child—because the trait is recessive, they will not exhibit the trait, but there is a 25% chance that they will have a child with the homozygous recessive genotype. To be sure, there are other ways of having such a child: a heterozygote can mate with an individual who is homozygous for the recessive allele (and then have a 50% chance of a child with the homozygous recessive genotype), or two homozygotes for the recessive allele can mate (and then have a

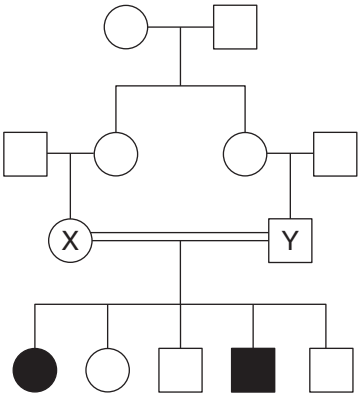


FIGURE 1.10

Pedigree illustrating autosomal recessive inheritance. The double horizontal lines indicate mating between people who are related (in this case, first cousins); the other symbols are explained in the legend to Figure 1.9.

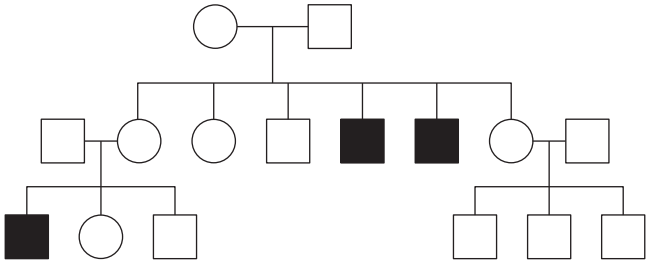


FIGURE 1.11

Pedigree illustrating sex-linked recessive inheritance. The symbols are explained in the legend to Figure 1.9.

100% chance of a child with the homozygous recessive genotype). For common traits, especially those which don't have any impact on reproduction, such matings are also common (as, e.g., with the O allele of the ABO blood groups). But if a trait is very rare, or very debilitating, then virtually all matings that produce children with the homozygous recessive genotype involve two heterozygotes, who thus do not exhibit the trait. For example, until very recently people afflicted with cystic fibrosis, which is an autosomal recessive disease, invariably died from the disease before having children. Thus, all children born with cystic fibrosis were born to people without the disease but who, therefore, are heterozygous for the allele causing the disease (barring new mutations). And what is the chance that a couple with one child with cystic fibrosis will have another child with cystic fibrosis? Hopefully, the answer is obvious to you by now: 25% (if not, do the Punnett square!).

Also note that a new symbol appears in the pedigree in Figure 1.10, and that is a double horizontal line between individuals X and Y. Further inspection reveals that individuals X and Y are related: they are first cousins, having one set of grandparents in common. The double horizontal line thus indicates a **consanguineous** marriage (one involving related individuals), which results in consanguinity or **inbreeding** in the children. Inbreeding will be discussed in more detail in Chapter 5; just realize for now that inbreeding results in an increase in homozygosity in the children. This happens because the same allele can be transmitted from one of the grandparents to both parents and then to both of their children (the first cousins). There is then a 25% chance that this same allele gets transmitted from both of the first cousins to their child. Overall, a child of first cousins has a 1/16 (or, about 6%) chance of being homozygous for an allele that was present in one of the grandparents of the first cousins. For a rare trait, this can be much higher than the chance of a homozygous recessive child from two unrelated parents. In fact, some extremely rare traits are known only from children of related parents, and in general, an increase in the frequency of related parents among children with a particular trait is an indication that the trait exhibits autosomal recessive inheritance.

Finally, consider the pedigree shown in Figure 1.11. Here, we see that individuals with the trait have parents who do not have the trait, which suggests recessive inheritance. However, only males have the trait. These are the hallmarks of an **X-linked recessive** trait, where the responsible gene is located on the X chromosome. A female who is heterozygous for an X-linked recessive trait will not exhibit the trait and is sometimes said to be a **carrier** for the trait. However, 50% of her sons will inherit an X chromosome with the recessive allele and hence will exhibit the trait. And, there is a 50% chance that her daughters will inherit an X chromosome with the recessive allele from her and hence also have a 50% chance of having a son with the trait. A famous example involving an X-linked recessive trait is that of Queen Victoria (1837–1901) of England, who bore three daughters who turned out to be carriers of hemophilia as well as a son with the disease. Several of her descendants married into various European royal families, resulting in numerous hemophiliacs among these royal families in succeeding generations.

Autosomal dominant, autosomal recessive, and X-linked recessive are the most common modes of inheritance of human traits. The other possible types of inheritance (X-linked dominant and Y-linked) are relatively rare and are left as exercises for you to work out (there is also **mitochondrial DNA**, which is maternally inherited, as discussed in Chapter 9). In working out how a trait is inherited from pedigrees, it is important to keep in mind that lots of families

(ideally, with lots of children) are needed to establish the mode of inheritance. Any individual family, especially if there are only a few children, may not be informative enough. For example, if two parents without a trait have a son with the trait and a daughter without the trait, this could be X-linked recessive inheritance, but it could also be autosomal recessive inheritance. If the trait is observed to occur only in male children in many families, then there would be conclusive evidence for X-linked inheritance.

WHAT IS—AND ISN'T—INHERITED

Take a look sometime at the unfortunately named Online Mendelian Inheritance in Man Web site (unfortunately named because Mendelian inheritance also applies to women!) accessible at http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim. This is a catalog of traits that exhibit, as the name suggests, Mendelian inheritance—that is, these are traits for which the variation is inherited in an autosomal/X-linked, dominant/recessive fashion. The variety of traits that exhibit Mendelian inheritance is truly staggering. The ability to roll one's tongue, attached versus free earlobes, wet versus dry ear wax, widow's peak (a pointed front hairline)—these are just a few of the traits that have been suggested to exhibit Mendelian inheritance. My own favorite is #108390, urinary excretion of the odoriferous component of asparagus, which simply means that after eating asparagus, some people have smelly urine and some people don't. Smelly urine is inherited as an autosomal dominant trait, although recent work suggests that in fact everyone has smelly urine after eating asparagus, and rather it is the ability to smell the smelly urine that varies among people and is inherited (you can look it up for the details).

Moreover, these are not the only traits that are inherited. Many traits have a more complex genetic basis and/or are influenced by both genes and the environment. Such traits include many that are of anthropological interest (such as variation in skin pigmentation, discussed in Chapter 20), as well as many common diseases (such as susceptibility to adult-onset diabetes or heart disease). These traits are generally known as **quantitative traits**, because the variation is continuous, meaning that the only limit on the values that the phenotype can take is the precision of the instrument used to make the measurement. For example, measure someone's height with a meterstick and you might get a value such as 183 cm (for the metrically challenged, this is about 6 ft). Use a laser and you might get a value like 183.241 cm. Another way to think about quantitative traits is that no matter how similar two phenotypes are, in theory it is always possible for someone to come along with a phenotype that is in between them (e.g., two people may be 183.241 and 183.242 cm in height but then a third person may be 183.2415 cm). In contrast to quantitative traits are **discrete traits**, which are usually either present or absent or exist in a few discrete categories that are counted as whole numbers (i.e., there are just four possible ABO blood group types). Quantitative traits are also influenced by the environment, whereas discrete traits generally depend only on the genotype (e.g., your ABO blood group genotype completely determines your ABO blood group type regardless of the environment, whereas your height is influenced by your genes, your diet, your overall health, etc.).

A simple example as to how the environment and the genotype interact to determine the phenotype is provided by a very rare type of deafness that is caused by both a particular mutation and an exposure to an antibiotic during childhood. If you have the "normal" genotype at this gene, you will have normal hearing. And, if you have the "deafness" genotype, but never take antibiotics, you will also have normal hearing. But, if you have the "deafness" genotype and you take an antibiotic during childhood (typically because you have some infectious disease, most commonly an ear infection), you will become deaf. It takes both the deafness genotype and the environmental exposure to an antibiotic to produce the deafness phenotype. For those of you who are parents, I hasten to add that this particular deafness mutation is *extremely* rare—it has only been found in a few families around the world—so you should not be concerned that you risk making your child deaf by administering antibiotics in case of an illness!

The analysis of quantitative traits gets very complicated very quickly and is beyond the scope of this book. But, since quantitative traits are also of great interest to people, it is important to know how to think about them. Let's take weight as an example. Suppose my weight is somewhat heavier than average, and I would then like to know how much of my excess weight is due to my genotype, and how much is due to what I eat and my level of physical activity (this would be my environment). If it turns out that my genotype is mostly responsible, good, then I can blame my parents for my excess weight, but if it turns out that my diet/exercise is mostly responsible, then I have only myself to blame. To figure this out, let's carry out the following thought experiments: create identical copies of me (i.e., clones with the exact same genotype at all genes as me) and put them on all possible diets and exercise regimens, and then see how often these clones have excess weight. At the same time, take everyone else, have them eat what I eat and exercise as much as I do, and see whether they also end up with excess weight or not. In the first experiment, we get an idea as to how my genotype "performs" in different

environments, while in the second experiment, we get an idea as to how much of an impact my own environment has when many different genotypes are exposed to it. If my clones tend to always have excess weight no matter the diet, then good, my genotype is to blame and I can eat whatever I want without feeling guilty. But if many different genotypes tend to have excess weight with my diet and level of exercise, then that would indicate that my environment is to blame for my excess weight (in which case, I will blame advertisers for enticing me to eat a poor diet!).

Obviously, we can't actually carry out such an experiment with humans but we can with other organisms. In particular, we can take cuttings from plants, thereby creating many different individual plants with identical genotypes and then raise the cuttings in different environments. An example where this was actually done is shown in Figure 1.12, where seven different cuttings (representing seven different genotypes) of a weed called *Achillea* were raised in three different environments (low, medium, and high altitude). Now, let's suppose I have bad news and good news for you. The bad news is that you have been very bad in this life, and so in your next life you will be reincarnated as an *Achillea* weed. The good news is that I will let you choose which genotype you can come back as. Look at Figure 1.12—which genotype would you choose? Your answer should be, well, it depends on which environment you end up in—the "best" genotype depends on whether you are planted at low, middle, or high altitude. Suppose I instead let you choose your environment—at which altitude would you like to be planted? Again, your answer should be that your choice of altitude depends on which genotype you come back as. The important take-home message: *there is no one genotype that performs best across all environments, and there is no one environment that is best for all genotypes*. To the extent that these sorts of experiments have been done, this is the usual result. Therefore, in order to understand how genes and environments interact to produce phenotypes, it is necessary to understand the **norm of reaction**—how phenotypes vary across different environments for different genotypes (as is shown in Figure 1.12 for a very limited number of genotypes and environments). Individuals who have a particular talent—academic, artistic, musical, athletic, and so forth—are often assumed to be innately talented, that is, that they would be talented regardless of the environment. Or, you may think that anyone raised in the same environment as a talented individual—given the same training, opportunities, circumstances, encouragement, experiences, and so forth—would develop a similarly exceptional talent. But the norm of reaction shows that both views are unfounded and most likely wrong: individuals who have an exceptional phenotype probably owe this to the combination of their particular genotype and their particular environment. Put the same genotype in a different environment, or expose a different genotype to the same environment, and you most likely won't get the same exceptional phenotype.

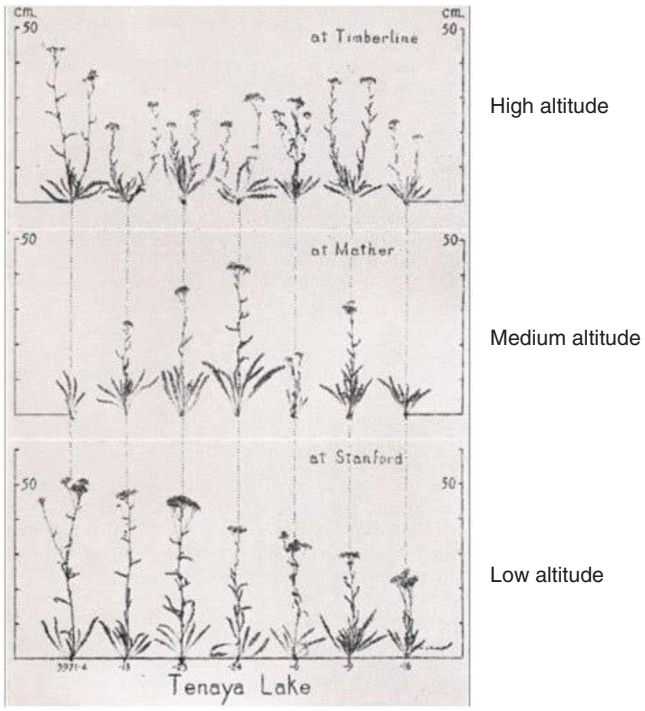


FIGURE 1.12

Image of cuttings from seven different *Achillea* plants grown at three different altitudes. The plants in each column are cuttings from the same plant and hence genetically identical. Modified with permission from Clausen, J., Keck, D.D., and Hiesey, W.H., "Experimental studies on the nature of species," *Environmental Responses of Climatic Races of Achillea, Volume 3: Carnegie Institute of Washington*, Washington, DC, 1948.

Finally, we have been concerned in this section with what sorts of traits are inherited, but it is also important to keep in mind that many human traits are *not* inherited. It is often thought that if a trait "runs in families," then it must be inherited. This is what I call the "fallacy of familiality"; there are many traits that tend to run in families but are not inherited. For example, family members tend to share political viewpoints and religious beliefs more often than people chosen at random, hence political viewpoints and religious beliefs are familial but they are not inherited. A particularly sobering example is pellagra, a disease due to vitamin deficiency that increased significantly in prevalence in the southern United States in the early 1900s. The actual cause for the increase in pellagra was poor nutrition associated with poverty, but a commission appointed to study the disease concluded

that it was instead inherited because it tended to run in families. Apparently, the commission did not realize that poverty also tends to run in families, and it took a long time before it was realized that simply improving the quality of the diet was sufficient to eliminate the disease.

CONCLUDING REMARKS

In this first chapter, we've covered the basics of how human traits are inherited. We've seen that genes are particulate and that alleles are not influenced by phenotypes—an O allele inherited from an AO parent behaves exactly the same as an O allele inherited from an OO parent. We've also seen that alleles at different genes are inherited independently—unless the genes in question are linked, that is, located close to one another on the same chromosome. We've gone through the properties of autosomal recessive, autosomal dominant, and sex-linked recessive inheritance, and how the mode of inheritance of a trait can be inferred from studying families. We've also briefly touched upon quantitative traits and distinguished what is inherited (passed on by genes) from what is merely familial. You now know (more or less) as much about genes as scientists did before they figured out what genes actually are, what they do, and how they do it, which is what we shall turn to next.

CHAPTER 2

WHAT GENES ARE, WHAT THEY DO, AND HOW THEY DO IT

In the preceding chapter, we saw how genes are inherited in humans without knowing anything about what genes actually are. Now we will go through the basics of molecular genetics, namely, what genes are made of, what genes do, and how they do it. This may seem backward—why not first present what genes are before discussing how they are inherited—but in fact, this order reflects history: all the mechanics of inheritance were worked out long before scientists were able to figure out what a gene is actually made of. And, as we shall see, knowing how genes are inherited provided crucial information for ultimately figuring out what they are made of.

CHROMOSOMES, PROTEINS, AND NUCLEIC ACIDS: FIGURING OUT WHAT GENES ARE

Mendel published his laws of inheritance, painstakingly worked out from his experiments on garden peas, in an obscure journal in 1866. His pioneering work remained unknown to the scientific field until it was independently replicated and then rediscovered in the early 1900s, well after his death (a prime example of the importance of properly publishing and publicizing one's results!). With the widespread acceptance of Mendel's laws, it was quickly realized that whatever genes are, they must: exist in pairs in cells; come apart (segregate) from one another during the production of eggs and sperm; come back together after fertilization to form individuals; and be contributed equally from the male and female parents. And, thanks to advances in microscopy that coincided with the rediscovery of Mendel's laws, it was just as quickly realized that there were already structures observed in cells that appeared to fulfill all of these requirements: namely, chromosomes, which as we saw in the last chapter exist in pairs in somatic (body) cells but singly in egg and sperm cells.

Thus, the first attempts to figure out what genes are made of focused on determining what chromosomes are made of. Chromosomes occur within a specific structure in the cell called the **nucleus**, while the rest of the cell contents are known as the **cytoplasm**. It turns out that chromosomes consist of two substances: **proteins** and **nucleic acids**. Proteins are important components of our bodies, and there are several different kinds of proteins: **structural proteins** provide support to cells, skin, hair, bones, and so forth, and enable muscles to contract and relax; another class of proteins called **enzymes** carry out all of the chemical reactions necessary to support life, such as **metabolism**; and other proteins, such as **receptors**, **hormones**, and **antibodies**, are involved in communication among cells and between our cells and the environment. Proteins (Figure 2.1) consist of one or more **polypeptides**, which are linear chains of up to several hundred **amino acids**, of which there are 20 major varieties and several minor ones. By contrast, nucleic acids have a much simpler structure. The primary nucleic acid, DNA (shown in Figure 2.2, along with another nucleic acid, RNA, that we'll get to soon), comprises just four **nucleotides** (adenine, cytosine, guanine, and thymine, conveniently abbreviated A, C, G, and T, respectively) that were originally (and quite mistakenly) thought to be repeated in blocks consisting of one of each nucleotide.

Another property that genes must fulfill is that they must be capable of existing in a large number of

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

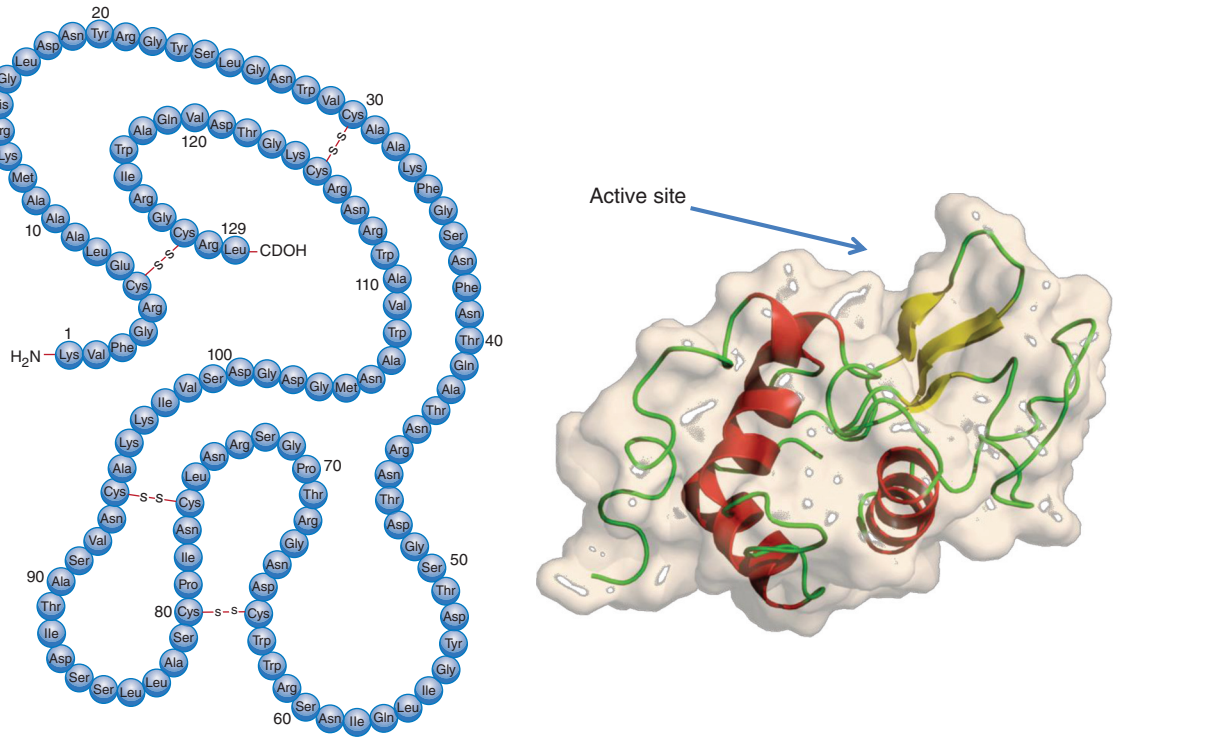


FIGURE 2.1

The structure of the lysozyme protein. Left, the linear amino acid sequence. Each circle shows one of the 129 amino acids (Lys = lysine, Val = valine, etc.) that constitute the active protein, which are arranged in a linear (polypeptide) chain. The chain has a beginning (the NH_2 end) and an end (the COOH end). There are also links called disulfide bridges (-S-S- bridges) between Cys (cystein) amino acids in the peptide chain, which assist the peptide chain into folding into the three-dimensional configuration shown on the right. The arrow points to the active site, which is where lysozyme binds to carbohydrates and cleaves them. Left: Modified with permission from John Kimball's Biology Pages (http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/L/Lysozyme.html; used with permission). Right: Modified with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Lysozyme.png).

different varieties. Given the erroneous view of the simple, repetitive nature of DNA, it did not seem possible for DNA to fulfill this requirement. The large number of different amino acids, on the contrary, means that the number of possible different polypeptides is vast indeed. If genes consist of linear chains of just 100 amino acids each, then there are 20^{100} different possible genes, or about 1.2×10^{130}. This number is so big that it is impossible for mere mortals to comprehend just how big it is—for comparison, the number of atoms in the universe has been estimated to be on the order of "only" 10^{80} or so! Proteins can, therefore, easily accommodate the unknown, but undoubtedly large, variety of genes that must exist, and so they became the focus of attention. Moreover, proteins fulfill a large number of different roles in the body, hence providing a means of linking genotype to phenotype: change the gene by changing the amino acid sequence of the protein and the phenotype would change, or so the thinking went (as we shall see in a minute, this is not so far removed from the truth). Thus, it is not so surprising that for many years proteins were thought to be the genetic material, and nucleic acids were largely ignored, as they were thought to merely provide structure to the chromosome.

The situation changed in 1944, with the publication of an experiment by Avery et al. (1944) that is imaginatively called the Avery–Macleod–McCarty experiment. Briefly, as shown in Figure 2.3, it had previously been shown that injecting a virulent (disease-causing) strain of *Streptococcus pneumoniae* bacteria into mice caused the mice to get sick and die, whereas injection of a nonvirulent strain of the bacteria had no effect on the mice. If the virulent bacteria were killed first with heat and then injected into mice, the mice survived. But mix heat-killed bacteria of the virulent strain with living bacteria of the nonvirulent strain and the mice injected with this mixture got sick and died. The obvious implication was that some factor was transferred from the killed virulent strain to the living nonvirulent strain, thereby altering the nonvirulent strain genetically and causing it to

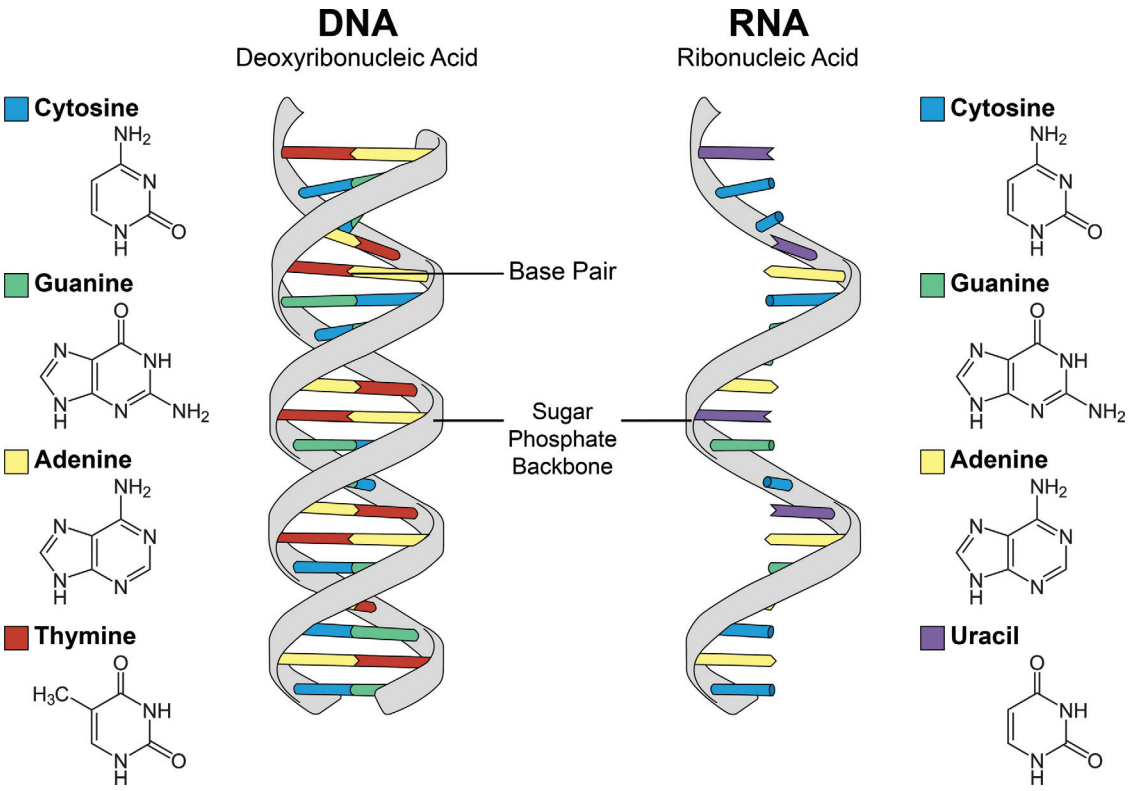


FIGURE 2.2

Structure and composition of DNA and RNA. Three of the nucleotides are identical between DNA and RNA, while thymine in DNA is replaced by uracil in RNA.

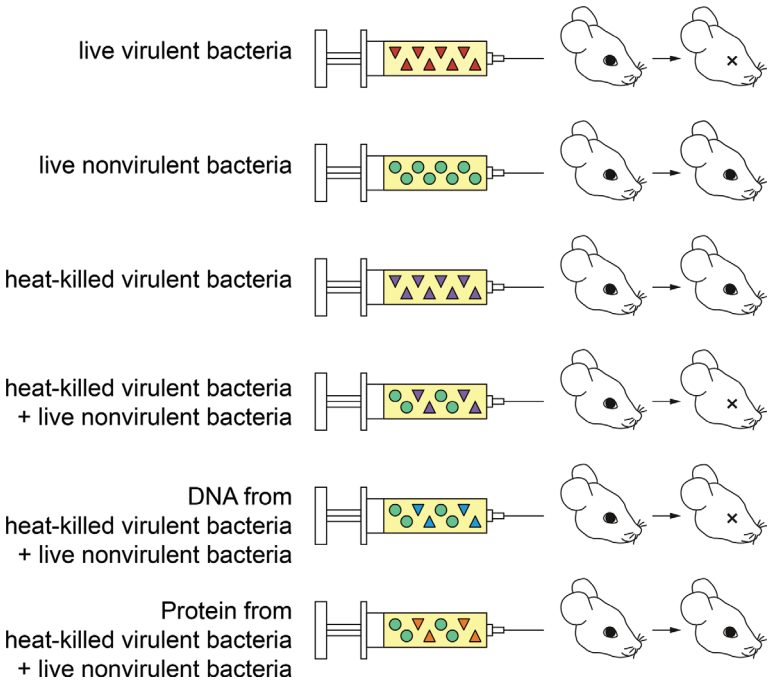


FIGURE 2.3

The results of the Avery–McCleod–McCarty experiment. See text for details; this experiment demonstrated that DNA and not protein is the disease-causing agent in bacteria.

become virulent. Avery, Macleod, and McCarty set out to determine whether this (genetic) factor was protein or DNA by studying chemical extracts that contained either DNA or proteins from the killed virulent strain. When they treated the extracts to destroy the proteins (but not the DNA), mixed the treated extracts with the living nonvirulent strain, and injected mice with this material, they ended up with dead mice. Evidently, the genetic factor was not protein. But when they treated the chemical extracts to destroy the DNA (but not the proteins) and mixed the treated extracts with the living nonvirulent strain, the injected mice showed no ill effects. The conclusion from this simple but elegant experiment was that DNA was transferred from the killed virulent strain to the living nonvirulent strain of bacteria, and the transferred DNA altered the nonvirulent strain genetically, causing it to become virulent. Thus DNA, not protein, was the stuff of genes.

As is so often the case in science, the results of this worldview-transforming experiment were greeted with extreme skepticism—so entrenched was the idea that genes must be made of proteins. Some claimed that the supposedly protein-free extracts that resulted in transformation of the nonvirulent strain to virulence must be contaminated with a small amount of protein. Others disputed the relevance of what happened in bacteria to the genetics of higher organisms such as humans, as bacteria had neither paired chromosomes nor sexual reproduction. Plus, many prominent scientists at the time were concerned with the function of genes, not their structure, and dismissed the debate as to whether protein or DNA was the substance of genes as being irrelevant to their interests. However, the Avery–MacLeod–McCarty experiment did inspire others to begin investigating DNA seriously, and the results were soon verified beyond any doubt. Moreover, with the determination of the structure of DNA by the classic work of James Watson and Francis Crick, what this structure could tell you about gene function became abundantly clear, as discussed later.

THE STRUCTURE OF GENES AND WHAT THEY DO: THE CENTRAL DOGMA AND THE FLOW OF INFORMATION

Working with data from another researcher, Rosalind Franklin, that they either did or did not have permission to view (depending on who is telling the story), in 1953, James Watson and Francis Crick worked out the well-known, iconic double helix structure of DNA (Watson and Crick 1953). A key insight was provided by the discovery of Erwin Chargaff that for the DNA of any cell or organism, the amount of nucleotide A always equaled the amount of nucleotide T, and the amount of nucleotide C always equaled the amount of nucleotide G (Chargaff et al. 1951). The DNA structure of Watson and Crick (1953) nicely accounts for this relationship of %A = %T and %C = %G, as the double helix consists of two intertwined single strands of nucleotides in which an A on one strand is always paired with a T on the other strand (and vice versa), while a C on one strand is always paired with a G on the other strand (and vice versa). Thus, given a DNA sequence of one strand—such as AGGCTAT—it is a trivial task to write the DNA sequence of the **complementary** (other) strand (in this case, TCCGATA).

Given the structure of DNA, further details as to what genes do and how they do it were quickly worked out. It turns out that those scientists who thought that genes were proteins were not so far off the mark after all, as genes contain the information that specifies the amino acid sequence of each polypeptide chain. Thus, proteins are the products of genes, and genes can be thought of as the blueprints, or set of instructions, that tell the cell the order of the amino acids for each polypeptide chain. As we shall see in the next section, there is an intermediate step in the process of making proteins that involves another nucleic acid, **ribonucleic acid** (**RNA**; Figure 2.2). Before getting into the details of how genes do what they do, however, there is an important point to consider and that is the flow of information in cells: the DNA sequence of a gene specifies the corresponding RNA sequence, which in turn specifies the corresponding amino acid sequence of a polypeptide chain. Information, therefore, flows in one direction: DNA → RNA → protein. DNA can make DNA, but RNA cannot make RNA and protein cannot make protein, nor can information flow backward. This concept of a one-way flow of information was developed by Francis Crick, and he called this idea the "Central Dogma" of molecular biology (Crick 1970). Later, Crick admitted that his definition of "dogma" differed from the usual dictionary definition; he thought that it meant a hypothesis based on little experimental evidence, whereas in fact it means a belief that cannot be doubted—hardly a good term for a scientific hypothesis! If he had known better, he would probably have called it the "Central Hypothesis" or the like. Anyway, the name stuck. Like any good dogma, the Central Dogma does have exceptions, the main one being that under some circumstances, RNA can be used to make DNA—indeed, some viruses make their living by converting their RNA to DNA. However, the fundamental idea of the Central Dogma, in which the DNA sequence of a gene specifies the corresponding RNA sequence, which in turn specifies the corresponding amino acid sequence, remains an extremely useful concept.

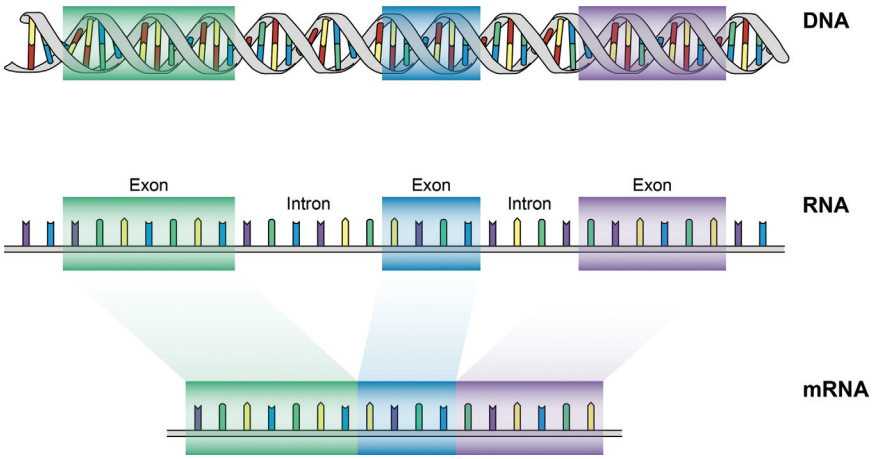


FIGURE 2.4

The intron–exon structure of genes. Exons are the protein-coding sequences, while introns are noncoding sequences within genes. Introns are transcribed into RNA and then spliced out to form the messenger RNA (mRNA). Modified with permission from Wikimedia Commons (http://commons.wikimedia.org/wiki/File:DNA_exons_introns.gif).

HOW GENES DO WHAT THEY DO: TRANSCRIPTION AND TRANSLATION

So how does the cell turn the DNA sequence of a gene into the amino acid sequence of a polypeptide chain? Consider first the structure of a gene, illustrated in Figure 2.4. The most surprising feature of this structure is that the typical gene has several different parts: the DNA sequence that corresponds to the amino acid sequence (the **coding sequence**), plus additional DNA sequence both before (**upstream sequence**) and after (**downstream sequence**) the coding sequence. Even more surprising, the coding sequence is interrupted in several places by **noncoding sequence** (that has no counterpart in the amino acid sequence). This **intron-exon** (intron for noncoding, intragenic regions, exon for coding regions) structure was discovered in 1977 (Berget et al. 1977; Chow et al. 1977) and came completely out of the blue—nobody was expecting that genes would come in bits and pieces.

As touched upon previously, it is the fundamental pairing of one nucleotide with another, usually referred to as **base-pairing** (viz., A with T and G with C) that leads to the transfer of information, both during **replication** (synthesis of new DNA) and **transcription** (synthesis of RNA, in this case **messenger RNA**, or mRNA). We will consider replication later in this chapter; for now let's see how transcription and **translation** (how a polypeptide chain is synthesized from mRNA) work.

RNA differs from DNA in several respects, most notably in that while it also consists of four nucleotide bases, three of these are the same as in DNA, but one is different: instead of T (thymine—see Figure 2.2), RNA has U (uracil). The important steps in transcription are shown in Figure 2.5. First, the DNA becomes single-stranded, and mRNA synthesis starts at a specific location (the **promoter**) in the upstream sequence. The mRNA is built sequentially, nucleotide base by nucleotide base, using the same base-pairing rules as with the DNA double helix, except that A in DNA specifies U in mRNA. So, given a DNA sequence, it is straightforward to figure out the corresponding mRNA sequence. The mRNA synthesis continues through the gene, including exons and introns, until it reaches **termination sites** that mark the end of the sequence to be transcribed.

The mRNA strand then separates from the DNA, but much still remains to be done before the mRNA can be used in the creation of a polypeptide chain. The mRNA has to be processed by adding a "cap" at the

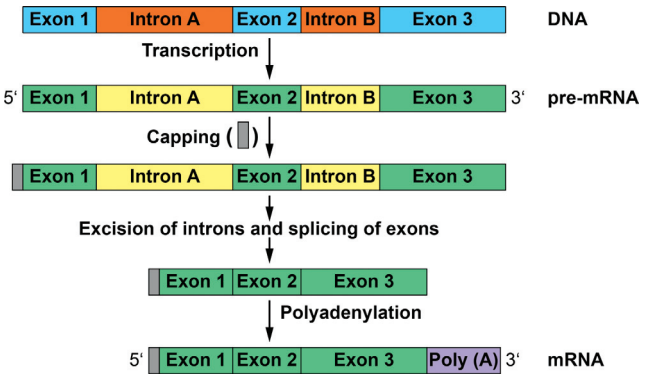


FIGURE 2.5

The steps involved in the production of mRNA from DNA.

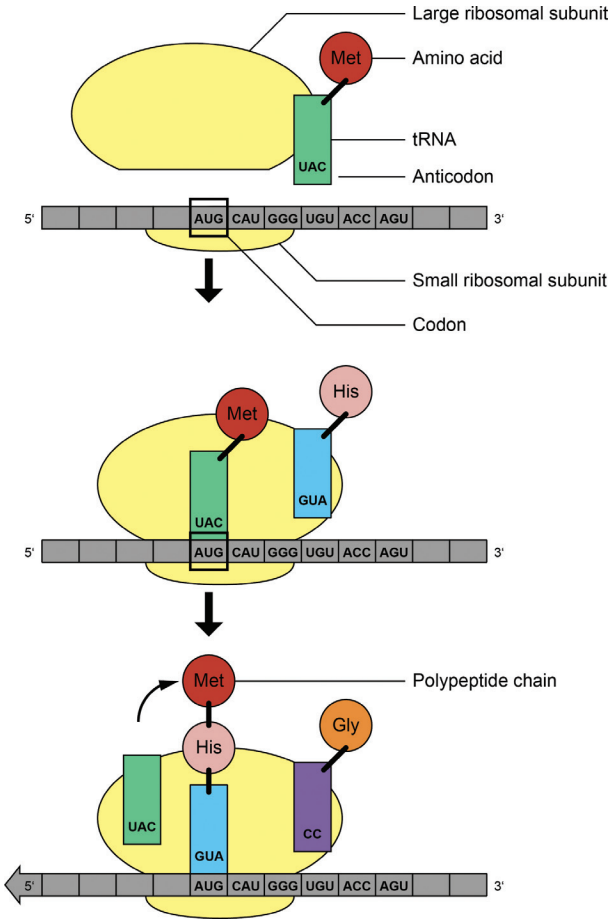


FIGURE 2.6

The steps involved in translation, which produces a polypeptide chain from an mRNA sequence. Transfer RNA (tRNA) molecules carry the corresponding amino acid and bind to the mRNA codon via the anticodon region of the tRNA; the amino acids are then assembled by the ribosome (which consists of large and small subunits).

head, followed by splicing out the introns and stitching together the introns. A "tail" consisting of multiple A nucleotides, known as a **poly-A tail**, is then added. Finally, the mRNA has to be transported from the cell nucleus into the cytoplasm, where translation of the mRNA into a polypeptide chain takes place.

Translation, shown in Figure 2.6, occurs on structures in the cytoplasm called **ribosomes**, which consist of around 80 different proteins plus a special kind of RNA called **ribosomal RNA** (rRNA). The process of making a polypeptide chain is quite complex and involves additional specialized RNA molecules called **transfer RNA** (tRNA). The tRNA molecules carry amino acids, and as the mRNA is "read" by the ribosome (much as a tape fed through a tape player), the corresponding tRNA brings the correct amino acid to the ribosome, which attaches the amino acid to the growing polypeptide chain. How the correct amino acid is actually determined is discussed later. After the polypeptide chain is finished, it detaches from the ribosome, is potentially subject to further chemical alterations (called **posttranslational modification)**, has to fold into the correct structure—which may involve forming chemical bonds with other polypeptide chains—and then has to be transported to the correct cellular location to function properly.

I've skipped over a lot of the details of transcription and translation as they are not really relevant for our purposes; anyone interested in the details can readily find them on the web or in any introductory genetics textbook. However, there are two important points to keep in mind. First, several hundred proteins are involved in transcription and translation, as well as special RNA molecules (rRNA and tRNA). These proteins include transcription factors that bind to DNA and initiate RNA synthesis, proteins to unwind the DNA double helix and make it single-stranded, RNA polymerases to make the RNA, proteins to cap and tail the mRNA and splice out the introns, proteins to transport the mRNA from the nucleus to the cytoplasm, and proteins that make up the ribosome. Where do these proteins and special RNA molecules come from? They are all encoded by the DNA—so the machinery for making proteins is also under genetic control. This means that the protein-making machinery is, like any other cellular process under genetic control, subject to being influenced by **mutations** (as discussed later in this chapter) and also has evolved according to the evolutionary principles discussed later on in this book.

Second, transcription and translation are amazingly complex processes, involving many steps (even more than those shown in Figures 2.5 and 2.6). Why such complexity? Are all those steps really necessary? Consider the following analogy: suppose in a university course in economics you are given the task of designing a plan for manufacturing a certain product, say doodads. You turn in to your professor the following plan:

1. The instructions for making doodads are kept in an office and are written in English, except that there are lots of additional words inserted before, after, and within the instructions that don't have anything to do with how to make doodads.
2. Secretaries copy the English instructions but translate them into German while doing so, and then all of the words that aren't actually part of the directions for making doodads are cut out and thrown away.
3. The German directions are then transported to the factory—which is in a different location—and the factory workers then follow the German directions in order to put together the doodads.

4. However, after a short period of time, the paper that the German directions are written on falls apart, so steps 2–3 have to be repeated continuously.
5. This same procedure is also used by the factory workers to make the machines needed to make doodads.

You can well imagine what sort of grade you would get for your plan! And yet, this is exactly what cells have to do in order to make proteins: DNA has to be transcribed into mRNA, which then has to be processed and transported from the nucleus to the cytoplasm in order to direct the synthesis of new polypeptide chains. Moreover, a general rule of thumb is that the more complex the process, the greater the chance that mistakes will occur. And this is certainly the case with transcription and translation; mistakes (in the form of mutations) have been documented in virtually all steps, sometimes leading to quite severe harmful effects.

So again, why such a complex process to make proteins, when it seems like your average university student could come up with a more efficient way of doing things? The answer is we don't know for sure, but we do have a pretty good guess and that has to do with controlling the amount of **gene expression** (i.e., how much mRNA—and corresponding polypeptide chain—is made from each gene). Gene expression is the critical aspect of growth and development. After all, the cells in all of the different organs of your body—skin, hair, muscle, bone, brain, heart, liver, kidneys, lungs, and so forth,—all have (more or less) the same DNA and the same genes arranged in the same order on the same chromosomes. If every organ has the same blueprint, then how do you end up with different organs? And how do you end up developing from a fetus to an adult, when your DNA stays (more or less) the same throughout your lifetime? The answer is that different genes are expressed in different organs and at different times as you grow and develop—basically, the heart pays attention only to those parts of the blueprint that pertain to making a heart and ignores all the rest (and exactly how that happens remains one of the big mysteries of life, although some aspects are beginning to be dimly understood). The same with growth—you have the same DNA and the same genes now as you did when you were born; the differences between you as an infant and you as you are now again reflect differences in gene expression. Moreover, as we shall see later, there is good evidence to think that changes in gene expression also played an important role in the evolution of our species.

So, the control of gene expression is a critical aspect of growth, development, and even evolution. And the numerous steps involved in gene expression provide numerous opportunities for fine-scale control: how much (and how fast) is the mRNA synthesized; how much (and how fast) is it capped, tailed, and spliced; how much (and how fast) is it transported from the nucleus to the cytoplasm; how much (and how fast) is it used to direct translation before it is degraded. Contrast this with the hypothetical situation of a cell that simply copies the DNA directly into a polypeptide chain; there would be much less opportunity for fine-scale control of gene expression.

As just one example of the potential importance of gene expression, consider the number of genes in humans versus other creatures. Before the sequencing of the human genome was accomplished in 2001, the number of genes in humans was unknown but estimated to be between 50,000 and 100,000. Fruit flies, by contrast, have about 14,000 genes, so for the many scientists who equated the complexity of an organism with the number of genes it has, this seemed like a comfortable distinction. To the surprise of many (and dismay of some), it turns out that humans have slightly less than 24,000 genes—that is, less than twice the amount of a fruit fly. Even the lowly flatworm has about 20,000 genes. Therefore, all the things that humans can do that fruit flies can't aren't just a product of humans having vastly more genes, because we don't (although it should be kept in mind that there are things fruit flies can do that humans can't, such as walk upside down on the ceiling). And as we shall see later, when we compare humans to our nearest living relatives, namely, chimpanzees, differences in genes alone don't seem adequate to explain all the phenotypic and behavioral differences between humans and chimps. Instead, it is differences in gene expression that probably account for a large part of the variation among different organisms.

Indeed, gene expression turns out to be even more complicated than simply how much mRNA a gene makes. It turns out that many genes have **alternative splice forms**, meaning that there are different ways of excising introns and stitching together exons from the transcribed mRNA. The result is that from the same gene one can get different mRNAs and different polypeptides (see Figure 2.7 for an example); variation in the amount of these different alternative splice forms is another form of variation in gene expression. And intriguingly, there is some (very preliminary) evidence to suggest that there is more alternative splicing in vertebrates in general (and, maybe, in humans in particular) than in other organisms—in fact, nearly all genes in humans have alternative splice forms. Thus, differences in organismal complexity may reflect variation in the complexity of gene expression, rather than differences in genes themselves. Or, in other words—and as in so much of life—what matters is not so much what you've got but what you do with what you've got.

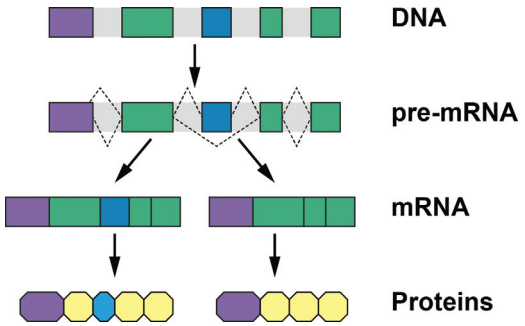


FIGURE 2.7

Alternative splicing, which is the production of different mRNA sequences (and hence, different polypeptide chains) from the same gene, due to the inclusion or exclusion of different exons (or parts of exons) in the mRNA.

THE GENETIC CODE

So far, we've skipped over a very important question: exactly how does the DNA sequence specify the amino acid sequence of the corresponding polypeptide? Let's consider the possibilities. Simplest would be that each nucleotide corresponds to a different amino acid, but the four different nucleotides in DNA aren't enough for the 20 different amino acids found in proteins. Suppose two consecutive nucleotides specify an amino acid: that gets us to 16 different possible pairs of nucleotides (4 × 4—write them out if this isn't obvious), but that is still not enough different combinations for 20 amino acids. But with three consecutive nucleotides per amino acid, there would be 64 possible combinations (4 × 4 × 4), more than sufficient for 20 amino acids. By this sort of reasoning, the **genetic code** (sequence of nucleotides that specifies each amino acid) must be at least triplet (i.e., at least three nucleotides per amino acid). Moreover, with 64 different triplets for only 20 different amino acids, the genetic code must be redundant: in principle, an amino acid could be specified by more than one triplet. A series of experiments in the 1960s demonstrated that the code is indeed triplet, worked out which amino acid corresponded to which codon, and verified that the code is indeed redundant.

The genetic code is depicted in Figure 2.8; the way to read this is that the nucleotide in the first position of the **codon** (sequence of three nucleotides) is along the left side (rows), the nucleotide in the second position of the codon is along the top (columns), and then the nucleotide in the third position of the codon is indicated in the box at the intersection of each row and column. The codon ACG, for example, specifies the amino acid threonine (thr). Note that three codons (UAA, UAG, and UGA) are **termination** (or "stop") **codons**, meaning that when the ribosome comes across one of these codons in the mRNA, translation stops and the polypeptide chain is finished. What about initiation of translation: what determines the position in the mRNA sequence where the ribosome starts making the polypeptide chain? Remember, mRNA typically contains untranslated regions both before and after the actual coding part of the mRNA, such as the cap and tail referred to previously (Figure 2.5), so translation does not just start at the very beginning of the mRNA. It turns out that the codon AUG, which encodes the amino acid methionine (met), is also used as the **initiation codon**. The first AUG in the mRNA sequence then marks where translation starts, so the first amino acid in every new polypeptide chain is methionine. However, this does not mean that all proteins start with methionine, because most polypeptide chains undergo further processing, including modification or cleavage of various amino acids, to form the active protein.

Note that the redundancy in the genetic code is not random; codons with the same nucleotides in the first two positions, but having A, C, U, or G in the third position, often code for the same amino acid (this is true for

Second position

First position (5′ end)	U	C	A	G	Third position (3′ end)
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys ACG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Amino acid names:

Ala = alanine	Gln = glutamine	Leu = leucine	Ser = serine
Arg = arginine	Glu = glutamate	Lys = lysine	Thr = threonine
Asn = asparagine	Gly = glycine	Met = methionine	Trp = tryptophan
Asp = aspartate	His = histidine	Phe = phenylalanine	Tyr = Tyrosine
Cys = cysteine	Ile = isoleucine	Pro = proline	Val = valine

FIGURE 2.8

The genetic code. Shown are the 64 possible triplets and the associated amino acid, indicated by the standard three-letter abbreviation. Stop indicates a stop codon.

8 of the 16 sets of codons with the same nucleotides in the first two positions; see Figure 2.8). Moreover, codons that differ only in the third position by U versus C, or by A versus G, often code for the same amino acid (this is true for an additional 6 of the 16 sets of codons). Thus, codons that differ only in the third position often code for the same amino acid. We will come back to this point later, when we discuss mutations.

DNA REPLICATION

Another property that genes must have is the ability to make copies of themselves that are nearly perfect, in order for genes to be transmitted to new cells that are made as the body grows and develops, as well as transmitted from parents to offspring. Why nearly perfect, and not perfect copies? Because there must be some mechanism by which new variation (i.e., mutations) can arise in genes, as otherwise there cannot be evolution (if this is not immediately obvious to you, wait for Chapter 5, where evolution will be discussed).

The process by which new copies of the genetic material are made is known as **DNA replication**. How this might happen was famously pointed out by Watson and Crick in their 1953 paper (Watson and Crick 1953) on the structure of DNA: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"—an understatement if there ever was one! Watson and Crick had noticed that since A always pairs with T (and vice versa), and C always pairs with G (and vice versa), the sequence of one strand of the DNA double helix tells you the sequence of the other strand. And indeed, what does basically happen during DNA replication is that the double helix unwinds, the two strands separate, and new strands are copied from the old ones, using the above base-pairing rules (Figure 2.9). Thus, DNA replication is **semiconservative**; the end product of one round of DNA replication is two DNA double helices, each consisting of one "old" strand and one newly synthesized strand.

THE CONSEQUENCES OF MUTATIONS

If a mutation—a change in the DNA sequence—arises, what effect will it have? (As to how new mutations actually occur, we'll get to that in the next section). The effect of the mutation depends on the type of mutation. Although there are many different kinds of mutations that can occur in DNA sequences, in this section, we will focus just on the changes that can occur at a single nucleotide position. Consider the following DNA sequence, and corresponding RNA and amino

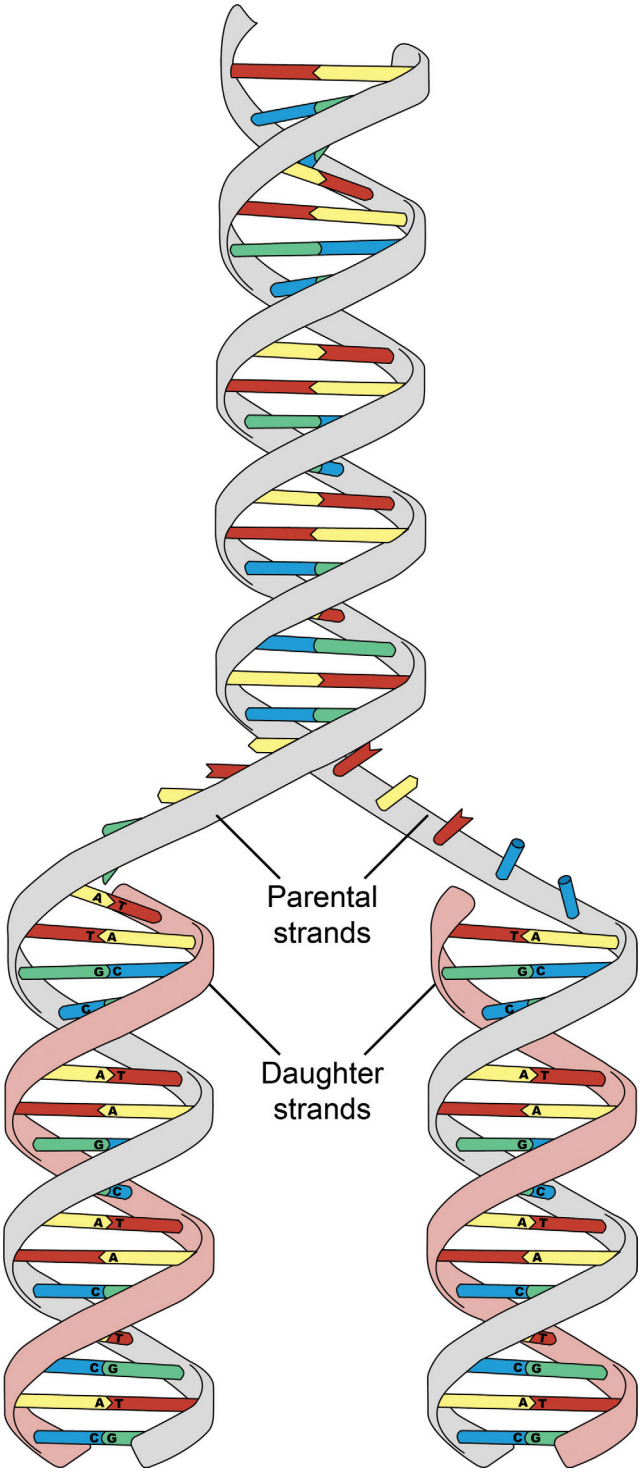


FIGURE 2.9

DNA replication, showing how parental (existing) strands "unwind" and then serve as the template for new (daughter) strands, using the base-pairing rules (A with T and G with C).

acid sequences, where each codon is separated by a dot (•):

DNA:	…ACG • AGG • CTC • AAT • CGG…
RNA:	…UGC • UCC • GAG • UUA • GCC…
Amino acids:	…cys • ser • glu • leu • ala…

Now, let's take just the first DNA codon (ACG) and see what happens when we change the third position (i.e., the G). Suppose this G is changed to an A. Then, the RNA sequence will change from UGC to UGU, because the DNA codon is now ACA. And the effect on the amino acid sequence? Look up UGU in the genetic code table in Figure 2.8, and you should see that UGU encodes cys (cysteine), the same amino acid that UGC encodes. So, this is an example of a **silent** or **synonymous mutation**: even though the DNA sequence has changed, the amino acid sequence of the corresponding polypeptide has not changed due to the redundancy in the genetic code.

Suppose the G in the ACG codon changes to a C. Then, the corresponding RNA sequence is UGG, which encodes a different amino acid, trp (tryptophan). So, one amino acid will be changed in the polypeptide sequence, while all the others remain the same. This is an example of a **missense** or **nonsynonymous mutation**: the change in the DNA sequence also changes the amino acid sequence of the polypeptide chain.

Suppose the G in the ACG codon changes to a T. Then, the corresponding RNA sequence is UGA, which turns out to be a stop codon. So now the amino acid sequence of the polypeptide chain will stop at this position. This is an example of a **nonsense mutation**: the change in the DNA sequence causes the polypeptide chain to terminate at this position, earlier than it should.

These are the three possible nucleotide substitutions that can occur at this position: G→A, G→C, and G→T. However, there is another type of change that could occur at this position and that is if the G is deleted from the sequence. What effect will this have on the amino acid sequence? Note that the cell's transcription and translation machinery has no way of "knowing" that a nucleotide has been deleted, so mRNA will be transcribed and translated exactly as if there never was a G in the DNA sequence. So the sequence of triplets—the **reading frame**—will be altered by the deletion of one base: the first position of the next codon will be read as the third position of the mutated codon, and so forth. In the present example, we have:

Original DNA:	…ACG • AGG • CTC • AAT • CGG…
Mutant DNA:	…ACA • GGC • TCA • ATC • GG…
Mutant RNA:	…UGU • CCG • AGU • UAG • CC…
Mutant polypeptide:	…cys • pro • ser • STOP
Original polypeptide:	…cys • ser • glu • leu • ala…

If you compare the mutant to the original polypeptide sequence, the first amino acid stays the same, but then the following two are different, and then there is a stop codon, which terminates translation of the mutant mRNA. This is an example of a **frameshift mutation**, which shifts the reading frame to compensate for the missing nucleotide. The end result is a very different amino acid sequence following the frameshift mutation, and often the polypeptide is shorter than usual, because a stop codon occurs within the new reading frame. However, sometimes a polypeptide that is longer than usual can result, especially if the frameshift mutation occurs near the end of the mRNA sequence, changes the original stop codon to an amino acid codon, and then a new stop codon occurs only in the (normally) untranslated region of the mRNA.

Note that any deletion (or insertion) of one or more nucleotide bases involving the reading frame is considered a frameshift mutation, unless three (or a multiple of three) nucleotides are deleted or inserted. When three (or a multiple of three) nucleotides are inserted or deleted, the reading frame is not altered. For example, in the DNA sequence we have been considering, suppose the entire AGG codon is deleted:

Original DNA:	…ACG • AGG • CTC • AAT • CGG…
Mutant DNA:	…ACG • CTC • AAT • CGG…
Mutant RNA:	…UGC • GAG • UUA • GCC…
Mutant polypeptide:	…cys • glu • leu • ala…
Original polypeptide:	…cys • ser • glu • leu • ala…

The result is the deletion of one amino acid. And what if the deletion of three nucleotides affects two codons? Suppose we take our original DNA and delete the three underlined nucleotides, which span two codons:

Original DNA:	…ACG • AGG • CTC • AAT • CGG…
Mutant DNA:	…ACG • AGG • CAT • CGG…
Mutant RNA:	…UGC • UCC • GUA • GCC…
Mutant polypeptide:	…cys • ser • val • ala…
Original polypeptide:	…cys • ser • glu • leu • ala…

The two codons CTC•AAT have been replaced by a single CAT codon, with the result that instead of two amino acids (glu • leu), there is now the single amino acid val (valine). So, overall there is a deletion of one amino acid and a substitution of another. However, sometimes there will only be a deletion of one amino acid, as the new codon will encode the same amino acid as one of the two codons that it replaces.

Of the various types of mutations that can occur at a single nucleotide position—silent, missense, nonsense, or frameshift—which do you think is likely to have the smallest effect on the function of a protein, and which would have the biggest effect? The general rule of thumb is that the more amino acids changed in

a polypeptide by a mutation, the bigger the expected effect on protein function. Hopefully, it is easy for you to see that silent mutations should have the smallest effect—in fact, they shouldn't have any effect on protein function, because by definition, they do not change the amino acid sequence of the polypeptide. Nonsense mutations, which cause premature termination of a polypeptide chain, should have a big effect on protein function, as should frameshift mutations, which alter many amino acids and often also cause the polypeptide chain to end before it should by creating a new stop codon. However, if the nonsense or frameshift mutation occurs very near the normal end of the mRNA, then sometimes there is little, if any, effect on protein function, because only a very few amino acids are changed.

You might think that missense mutations, which change only one amino acid, would not have a particularly big effect on protein function—and in most cases, you would be right, but in some cases, you would be quite wrong. For example, the β-globin polypeptide (part of the hemoglobin protein) has 146 amino acids; change the sixth one from glu to val and you get the disease sickle-cell anemia (discussed in more detail in Chapter 5). The polypeptide encoded by a gene called *CFTR* is involved in transporting ions across cell membranes and is quite big, with 1480 amino acids; deleting just one of these 1480 amino acids, at position 508, results in the disease cystic fibrosis (mentioned in Chapter 1). So, even seemingly innocuous changes at the DNA level can sometimes have profound effects on protein function.

WHAT CAUSES MUTATIONS?

There are two important sources of new mutations: mistakes during DNA replication and DNA damage. It turns out that DNA replication is an incredibly faithful process: the error rate for incorporating the wrong nucleotide during replication (i.e., the mutation rate) is about 1 mistake in every 100 million bases (written as 1×10^{-8}). If you don't think this is so great, try copying a sequence of 100 million As, Cs, Gs, and Ts by hand, and see how many mistakes you make! The chemical bonds formed when A pairs with T, and C pairs with G, help to ensure that the correct base is inserted into the newly synthesized DNA strand (Figure 2.9). Moreover, the enzyme which is primarily involved in synthesizing new DNA, known as **DNA polymerase**, has a "proofreading" ability: after inserting a base, the DNA polymerase can "check" if the base-pairing is correct. If not, the incorrect base is removed and a new base inserted. Without this proofreading, the error rate during DNA replication would be much higher, about 1 mistake every 10,000 bases (or 1×10^{-4}), which is an excellent example of the importance of proofreading your work!

In fact, RNA polymerase does *not* have any proofreading ability and hence does make roughly 1 mistake in every 10,000 bases in newly synthesized RNA. Why should RNA polymerase have a much higher error rate than DNA polymerase? While we do not know for sure, a likely reason is because each mRNA molecule lasts long enough only to make a few polypeptide chains. Thus, if a particular mRNA molecule does have an error introduced during transcription, even one that has a large impact on protein function, it will not have a correspondingly large impact on the organism because it will affect only a few protein molecules and only for a short time. By contrast, an error introduced during DNA replication will affect all of the mRNA molecules transcribed from that DNA and hence have a potentially much larger impact on the organism. It's like the difference between a mistake made by workers during the assembly of a car that affects only one particular car versus a mistake in design that results in the recall of all cars made from that design. The former is a lot less important than the latter (unless you happen to be the unfortunate individual who purchases the rare defective car, of course). You want to be sure that your design instructions (corresponding to DNA) are as error-free as possible—although, as we shall see later, in the case of DNA (although not in the case of car design), there is good reason to want some errors to occur.

The second important source of mutations is DNA damage. We are constantly exposed to chemicals and radiation from the environment that damage our DNA by modifying it in such a way as to make it difficult, if not impossible, to accurately replicate the DNA. In response, we have evolved sophisticated mechanisms to recognize and repair DNA damage, thereby minimizing its harmful effects. In fact, without these repair mechanisms, we would not be able to survive. As just one example, consider the case of a disease called **xeroderma pigmentosum** (XP). When ultraviolet light—even the amount that occurs naturally in sunlight, not just from tanning salons or other artificial sources—strikes DNA, it induces the formation of **thymine–thymine dimers**: at places where T is followed by T on the same DNA strand, the ultraviolet light causes an unusual chemical bond to form between the adjacent T (thymine) bases, rather than the usual base-pairing between each T and the A on the other strand (Figure 2.10). Normally when this happens, a specific enzyme chops the unusual thymine–thymine dimer out of the DNA and then it is replaced (via DNA polymerase and other enzymes) with Ts correctly base-paired with the As on the opposite strand. However, in XP, this process of chopping out and repairing the damaged bases (called **nucleotide**

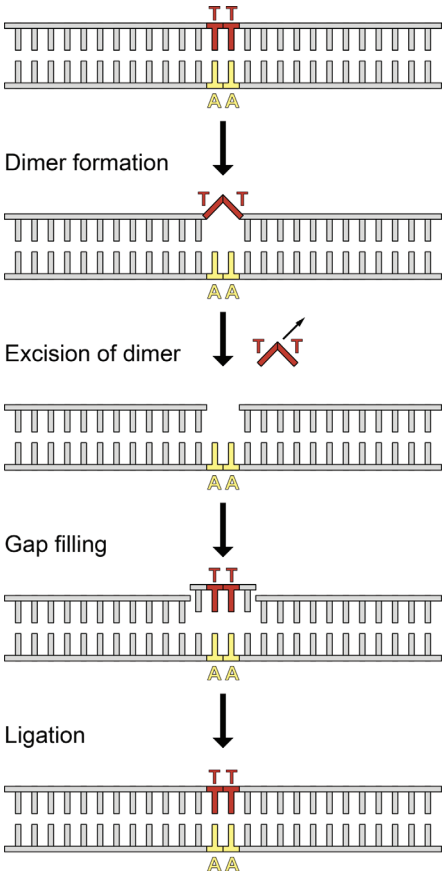


FIGURE 2.10

The formation of thymine dimers in DNA via exposure to ultraviolet light and removal via excision repair.

excision repair) is disrupted by a mutation that reduces or destroys the function of one of the enzymes involved. The severity of the symptoms of XP depends on the specific mutation, but in the worst cases, individuals suffer from severe freckling, sunburn, and greatly increased risk of skin cancer, even with very limited exposure to sunlight. The only "treatment" consists of limiting exposure to sunlight, meaning that unfortunate individuals with this disease must avoid sunlight at all costs. Even so, the majority of people with XP die before reaching the age of 20 years.

This sobering example well illustrates the importance of DNA repair mechanisms for sustaining life. The absence of just this one type of repair—nucleotide excision repair—is usually fatal. In fact, there are several types of DNA damage that can arise from exposure to various chemicals (including natural by-products of metabolism such as **free radicals**) and/or radiation. For all of these, we have evolved mechanisms to recognize and repair the damage—and a good thing it is, because as the case of XP amply illustrates, in the absence of DNA repair mechanisms, life as we know it could not survive on this planet.

A FINAL CAUTIONARY NOTE

The reader should be warned that I have greatly oversimplified the topic of this chapter, namely, molecular genetics. To some extent, of course, this is true of all of the chapters in this book—after all, this is an introductory textbook—but there is much more oversimplification in this chapter than in any other. This is partly because for our purposes, it is not really necessary to know all of the details of replication, transcription, and translation; moreover, if you are interested in the details, then you can readily find them in molecular genetics textbooks or on the Internet. But it is also partly because many of the details are only dimly understood, and new features are constantly being identified, even as I write this. I have already stressed the complexity of the processes involved in replication, transcription, and translation, but it is worth making the point again: our genome is a vast and mysterious place, with many untold wonders yet to be discovered.

CHAPTER 3

GENES IN POPULATIONS

Having covered how genes are inherited, as well as what they are, what they do, and how they do it, we are now ready to tackle how genes behave in populations. In case you are wondering, understanding how genes behave in populations is necessary for understanding how evolution influences genes, which in turn is important for trying to infer what happened in the past from patterns of genetic variation in populations living today—which is, to remind you, a major goal of molecular anthropology. So, in this chapter we will cover some important concepts about populations, while the next two chapters will discuss how genes behave in populations, and the various evolutionary forces and how they influence genetic variation within populations and genetic differences between populations. From an historical perspective, all of this information could have come before the preceding chapter on molecular genetics, because all of the ideas and concepts about genes in populations and how their variation is influenced by evolutionary forces were worked out long before anyone knew what a gene actually was. However, knowing something about what genes are does aid in understanding the ideas and concepts presented in this and subsequent chapters.

WHAT IS A POPULATION?

As we shall see in later chapters, many of the analyses that molecular anthropologists carry out are performed on data collected from populations. A critical aspect of these analyses, that often does not get enough attention, is determining who belongs to which population. So, let's start by defining what we mean by a population. Here is the scientific definition of the term "population": a spatial–temporal group of interbreeding individuals who share a common gene pool. This is a great example of how scientists like to do things: take a word for which everyone has at least some idea of the meaning (like "population") and define it with terms that nobody ever uses! But let's break down the terms and see what is going on. First, what is meant by "spatial–temporal"? Simply that a population occupies a particular geographic area (the spatial) and does so over a (relatively) long period of time (the temporal). So, assuming that you are reading this book because it is an assigned text for a particular class, we wouldn't consider you and your classmates to be a population, because even though you assemble regularly at one specific location (viz., the classroom or lecture hall), after class is over you disperse to other locations.

But simply being in the same location over a long period of time isn't enough to be considered a population; the individuals must **interbreed** (have children). So, we wouldn't consider you and your cat or dog to be part of the same population, even though you and your pet may be in the same location over a long period of time. Moreover, this interbreeding must have gone on for a sufficiently long enough period of time (i.e., several generations) that the people involved tend to share more alleles in common with each other than they do with people who belong to a different population. The collection of alleles that a population has is referred to as the **gene pool**, and population geneticists tend to think in terms of the gene pool rather than the individuals themselves that make up a population. This is because the gene pool is what is transmitted from generation to generation and thus survives beyond the life span of the people who actually make up the population at any given point in time. Don't worry if this definition of the gene pool seems rather vague, because it is, but by the end of the next chapter you should have a greater appreciation for the gene pool concept. So, to go back to our definition of a population, the people who make up a population today are descended from people who were also part of the same population (i.e., living in the same area, interbreeding, and sharing the same gene pool). So there you have it: a population is a group of people who occupy a specific location and have children together,

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

and whose ancestors have been doing so for many generations.

So far, so good—but there is a lot of imprecision in this definition, which then makes it difficult in practice to decide who is part of a specific population and who isn't. In particular, where do we draw the boundaries between human populations? In molecular anthropology, geographic or linguistic boundaries are often used to define populations. Thus, you may find studies of Africans, Europeans, and Asians, or studies of Yakuts, Evens, and Evenks (Yakut, Even, and Evenki are different languages spoken in Siberia). Sometimes you will come across the term **ethnolinguistic group**, which means that both cultural and linguistic criteria are used to define the groups studied—although very often, it is really just the language that is used to define the group, as often the name of the language and the name of the group are the same (as in the aforementioned Siberian example).

Still, how to define the groups studied remains a troublesome issue. Take me, for example. I was born in the United States but now live in Germany. For a molecular anthropological study, I would not be considered a native German, nor would I be considered a Native American, as in molecular anthropology that term is reserved for descendants of the prehistoric migration(s) to the Americas from Asia. The term that would most likely be used to describe me is "European–American," meaning an American of European ancestry (for those of you who would use the term "Caucasian," we'll see later why that is not such a good term), which while accurate is not very precise because Europe is a pretty big place. As it turns out much of my ancestry is from Germany, so maybe German–American would be more accurate. But most studies rely on self-described labels—that is, you ask people to tell you what group they belong to—and German–American is not a label that I would use to describe myself, I'd probably call myself a transplanted American or some such thing.

Another issue that often arises is that groups can be defined at different levels, and when groups defined at different levels are used in the same study, this can have an impact on the results and their subsequent interpretation. For example, one of the first international, collaborative projects to systematically study human genetic variation is the **International HapMap Project** (HapMap is short for "haplotype mapping"; we'll discuss haplotypes in Chapter 9), which in the first phase (The International HapMap Consortium 2003) collected and analyzed DNA samples from four populations: Yoruba from Ibadan, Nigeria; Japanese from Tokyo; Han Chinese from Beijing, China; and residents of Utah with European ancestry (primarily from northern and western Europe). The genetic data have been made publicly available, and it is a very useful resource that has been used in many studies. And yet, for the purposes of comparison, Yoruba is a language with some 20–40 million speakers in Nigeria; there are about 127 million Japanese and 1.2 billion Han Chinese; and your guess is as good as mine as to what to make of "Utah residents with European ancestry." Clearly it would not be a good idea to treat the genetic variation in these four individual populations as representative of all of Africa, Asia, and Europe. To their credit, the International HapMap Project is careful to emphasize this point, and yet there are studies that have done exactly that, either implicitly or explicitly. We'll see later on (in Chapter 11) an example of how the definition and sampling of populations can influence the outcome of a study. Fortunately, as we shall also see later in Chapter 11, there are new methods that allow the genetic data from individuals, rather than populations, to be analyzed. These analyses thus eliminate the need to group individuals into populations (although how the individuals were sampled is still an important issue). Still, these individual-level analyses cannot provide everything we might want to know about population history, and much of population genetic analysis focuses on the analysis of populations (which, after all, is why it is called population genetics!). So, you should always take care to notice how the populations in a study were defined and sampled.

THE CONCEPT OF "EFFECTIVE POPULATION SIZE"

When it comes to evolutionary genetics, not only is the definition of a population important but the size of the population also plays a key role in many concepts. You might think, well, what's the big deal, just count the number of people in the population and call it a day. However, as with the difficulties in defining the population, determining the size of the population is not as easy as it might appear. This is because in an evolutionary sense, what matters is not how many people are in the population but the size of the gene pool (i.e., how many different kinds of alleles there are and their frequency). This is because people die, but their genes live on in their descendants, and so evolution is concerned only about the gene pool and what you contribute to it and not about you as an individual (sorry to be the one to break this to you). And since alleles are transmitted from generation to generation via the production of children, it stands to reason that the size of the gene pool in the next generation is not related to the *total* population size but rather to the size of the group of people who are or will be having children—not everybody in a population has children. If we just count the variety of alleles, a population consisting of 100 people who never have any children would have

a bigger gene pool now than a population consisting of 10 people who do have children, but from an evolutionary standpoint, the latter population is actually larger than the first population, because the first population will not leave any descendants.

So the number of people who are or will be having children is an important aspect of figuring out how big a population is, but it is not the only one. Another important concept is that of the **ideal population**. To a population geneticist, an ideal population is one in which there are equal numbers of males and females, all individuals are unrelated, and everybody has the same chance of having children. Note that this is not the same as saying that everybody in fact has the same number of children; it just means that the probability that any particular individual has a child is the same for everyone in the population. If this is not clear, suppose in a population everyone gets one chance to have a child, and whether or not you have a child is determined by flipping a coin: heads, you have a child; tails, you don't. Then (assuming a fair coin) everyone has the same chance of having a child, but about 50% actually have children and 50% do not.

An ideal population is thus an example of a **biological model** (a simplified description of some aspect of the real world), and like any biological model, it has two properties: it enables predictions to be made about some aspect of the real world; and it is highly unrealistic precisely because it is a simplified version of reality. You may wonder about the value of predictions derived from unrealistic models, but as we shall see, sometimes even simplified versions of reality can be approximately correct. Moreover, a useful approach to investigating the real world is to start with a simple model, see what predictions results and then see what happens to those predictions when the model is made more realistic (and hence more complex); this is the approach we will use in the next two chapters.

To repeat, in an ideal population everyone is unrelated and has the same chance of having a child. This means that in an ideal population, every individual has the same chance of producing a gamete (egg or sperm) that ends up uniting with another gamete to produce a child, which in turn means that the two gametes that come together to make up a child have the same chance of coming from any two individuals. They could even both come from the same individual (remember, an ideal population is not supposed to be realistic!). And since the gametes carry the genetic material, thereby transmitting alleles to the next generation, the different kinds and frequencies of alleles in the gametes determine the gene pool in the next generation.

But if we now think about what happens in reality, some people have a much higher chance of having children than others, so they contribute more alleles to the next generation. Moreover, some people are related, and related people will have more alleles in common than unrelated people. For these two reasons, there is less genetic variation among the gametes than we might expect, given the actual size of the population. So, in the typical population, the size of the gene pool that is transmitted to the next generation is usually less than would be predicted if the population was really an ideal population. This brings us to the concept of **effective population size**: the size of an ideal population with a gene pool that would correspond to that of the real population, in terms of the variety and frequency of alleles in the gene pool. If we have a population of 100 individuals, but their gene pool is more like what we would expect to find in an *ideal* population of 50 individuals (i.e., 25 males and 25 females, all unrelated and all with the same chance of having children), then the *effective* population size (abbreviated as N_e) would be 50. So, we have used our biological model of an ideal population to interpret data from an *actual* population of 100 people.

As we shall see in more detail later, in population and evolutionary genetics, we are almost always concerned with N_e, rather than the simple census size of a population. Hopefully, this makes sense: in terms of evolution, it is the gene pool, not the bodies who carry the gene pool, that matters. Later on in this book (specifically, Chapter 5), we will see how we can use genetic variation data to estimate N_e. In the remainder of this chapter, we will examine how relaxing different assumptions about an ideal population can influence N_e.

THE SEX RATIO AND N_e

One assumption of an ideal population is that there are equal numbers of males and females; what happens if this assumption is violated? It turns out that we can write an equation that relates N_e to the number of females (N_f) and males (N_m) in the population:

$$N_e = 4N_fN_m/(N_f + N_m)$$

There are people who can look at an equation like this and immediately understand what it means; for the rest of us mere mortals, the best thing to do when confronted with an equation like this is to plug in some values for N_f and N_m and see what happens. Let's start with easy cases; what happens if there are 100 females and no males in the population? Then N_e turns out to be 0, and the same thing happens if there are 100 males and no females in the population. Hopefully that makes sense: if your population has only members of one sex, then as far as evolution is concerned that population is dead in the water because there is no possibility of producing children.

TABLE 3.1 ■ Values of N and N_e for different values of N_f and N_m

N_f	N_m	N	N_e
40	40	80	80
70	10	80	35
700	10	710	39
70	1	71	~4
700	1	701	~4

Suppose we have equal numbers of females and males. Then $N_f = N_m = N/2$, where N is the census size of the population. Plug this into the equation, do the algebra, and you should get $N_e = N$. Again, hopefully that makes sense: this equation is supposed to tell us what happens when we don't have an ideal population because we have unequal numbers of each sex. If in fact we do have equal numbers of each sex, then we do have an ideal population and hence N should equal N_e.

Table 3.1 shows what happens for some other values of N_f and N_m; let's go through these. If we have 40 males and 40 females, then $N_e = N = 80$, as we've already seen. But suppose we have 70 females and just 10 males. Even though N is still 80, N_e is only about 35, less than half of N. The fact that only 10 males are contributing to the next generation greatly reduces the gene pool, since they must contribute half of the genes. Even if all 70 females also contribute offspring, with only 10 males the number of males is the limiting factor when it comes to passing on genetic variation. Suppose we increase the number of females from 70 to 700 and keep the number of males at 10; then $N = 710$ but N_e goes only up to about 39. Suppose we go back to having 70 females in our population, but just one (happy!) male: N_e is only about 4. And if we have 700 females and just one (happy, but tired!) male, N_e is still only about 4. The take-home message: with unequal numbers of each sex, N_e is dominated by the limiting sex (i.e., the one with the smallest number), because half of the gene pool in each generation comes from each sex. So, if you have only one male or one female, it doesn't really matter how many you have of the other sex, because that one individual contributes half of the gene pool.

INBREEDING AND N_e

Another assumption of an ideal population is that all of the individuals are unrelated. Let's see what happens to N_e when we relax this assumption and allow for related individuals. As we saw in Chapter 1 in the example of autosomal recessive inheritance and first cousin matings, when individuals are related, this means that they share some alleles that they inherited from the same ancestor in a previous generation. If related individuals mate and have children, then this reduces the variation in the gene pool, because alleles that are **identical by descent** (i.e., identical because they are descended from the same ancestor) become overrepresented in the gene pool.

The way we measure the relatedness of individuals is with the **inbreeding coefficient**, which turns out to also measure the probability that the two alleles in an individual are identical by descent. The inbreeding coefficient is designated as **F** and can take values from 0 to 1. An inbreeding coefficient of 0 means that there is no inbreeding, while an inbreeding coefficient of 1 corresponds to mating with oneself—not something that we have to worry about in humans, of course, but this does occur in some organisms, such as some plants.

The formula for relating N_e to F is:

$$N_e = N/(1 + F)$$

As we did with the other formulas, let's try some values of N (the census size of the population) and F and see what happens to N_e. First, the trivial cases: suppose $F = 0$, corresponding to no inbreeding. Then $N_e = N$, which is as it should be—the formula indicates how N is reduced with inbreeding, so if there is no inbreeding, there is no reduction in N. Suppose $F = 1$ (complete inbreeding, from self-mating): then $N_e = N/2$. So the biggest reduction in the population size that we can get from inbreeding is by a factor of 2.

We will discuss inbreeding in humans in more detail in Chapter 5, but the largest values of F that have ever been observed in human populations approach 0.10 or so. With $N = 100$ and $F = 0.10$, $N_e = 91$. Most human populations that are considered highly inbred, such as the Pennsylvania Amish, have $F = 0.05$ or less (for comparison, $F = 0.0625$ for first cousin matings, like that shown back in Figure 1.10), which for $N = 100$ translates into $N_e = 95$. So the take-home message here is that while inbreeding can have detrimental consequences, the levels of inbreeding observed in typical human populations (usually $F < 0.01$) have a negligible effect on N_e.

VARIATION IN POPULATION SIZE OVER TIME AND N_e

A third assumption of an ideal population is that it remains constant in size from generation to generation; what happens if the population size actually fluctuates over time? Let's start by assuming that we have **discrete generations**, meaning that everyone is born at one time, reproduces at a later time, and then dies

after reproducing. While this is hardly a good model for humans, it is reasonably accurate for some organisms, such as salmon. We could assume **overlapping generations** instead (which would be appropriate for humans), but as it turns out, we would come to exactly the same conclusions as with discrete generations, only the math would be a lot harder. So for simplicity, we'll stick with discrete generations.

Start by assuming that we have a time span of t generations that we are interested in; we can then designate the effective population size for each generation as $N_1, N_2, \ldots, N_t$. It might seem that to get the average N_e over time, the obvious thing to do would be to simply take the average of the effective population size for each generation, that is, take the sum of $N_1 + N_2 + \ldots + N_t$ and divide by t. However, this is the **arithmetic mean**, and it is not the only way to calculate the mean value of a bunch of numbers. Another example of a mean value is the **harmonic mean**, which involves summing the reciprocal values of the observations, and it turns out that this is the correct way to estimate N_e when the population size fluctuates over time:

$$1/N_e = 1/t[1/N_1 + 1/N_2 + \ldots + 1/N_t]$$

So how does the harmonic mean compare to the arithmetic mean, and how is N_e influenced by changes in population size over time? Let's set $t = 10$ generations, and first suppose that the population size does *not* change over time. Set $N_1 = N_2 = \ldots = N_{10}$ and do the algebra (if the algebra is too hard, put in some arbitrary value for N, such as 100, and do the arithmetic); the answer you should get is that $N_e = N$, which hopefully makes sense to you: the above equation tells us what happens when the population size fluctuates over time, so when it does not fluctuate over time, there is no effect on N_e. Let's keep $t = 10$ generations, and set $N_1 = 50$ and $N_2 = N_3 = \ldots = N_{10} = 100$. Then, the arithmetic mean turns out to be 95, while the harmonic mean—which is our estimate of N_e—is 91, which is slightly smaller. Now suppose that $N_1 = 10$ and keep $N_2 = N_3 = \ldots = N_{10} = 100$; the arithmetic mean is 91, but N_e (the harmonic mean) is much lower, about 53. The take-home message: when the population size changes over time, N_e is dominated by small population sizes. Hopefully this makes sense; when N_e is small, the gene pool gets smaller because you have fewer alleles represented, and even though the population may grow later, it is not going to regain the alleles that were lost when it was small in size. So even though the current population size may be large, the size of the gene pool—and hence, our estimate of N_e—corresponds to when the population was much smaller.

DIFFERENTIAL FERTILITY AND N_e

The final factor to consider in our discussion of N_e is variation in family size among couples, also known as **differential fertility**. In an ideal population, everybody has the same probability of having children; what happens if we instead have big differences among couples in the number of children that they have? So far we have been concerned with factors that decrease N_e relative to N, but one interesting (albeit highly unrealistic!) potential consequence of differential fertility is that N_e can be bigger than N!

To understand how differential fertility influences N_e, we need to introduce another mathematical concept, namely, the **variance**. We have already introduced the mean (either arithmetic or harmonic) as one way to characterize a distribution of numbers, such as the population size for each generation, measured over some number of generations. So, if we are interested in the distribution of the number of children per couple, we can readily calculate the average number of children per couple, which would then be the arithmetic mean. But means aren't the only useful way to characterize a distribution of numbers; for some applications, we also want to know how close the individual observations are to the average value. For example, suppose I have six couples, of which two have one child each, two have two children each, and two have three children each. Then the average number of children per couple is $(1 + 1 + 2 + 2 + 3 + 3)/6 = 2$. Now suppose I have another six couples, of which five couples have no children and one couple has 12 children. Do the math; the average number of children per couple is also 2. Yet, clearly these are very different distributions; in the first case, most of the observations are close to or equal to the average, while in the second case, the observations are very different from the average. So we if want to know how close the individual observations are to the average—in other words, how much information the average actually provides about the distribution—we need a measure of how close each observation is to the average value. And that is where the variance comes in; the variance is obtained by taking each observation (i.e., the number of children for each couple), subtracting the average from it, squaring this number, and then summing these values across all observations. Or, when written out as an equation:

$$\text{Variance} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where x_i is each observation and $\bar{x}$ is the average of the n observations. So in the first example previously, the variance is 0.8, while in the second case, the variance is 24. The smaller the variance, the more similar the individual observations are to the average value.

To see how N_e is influenced by differential fertility, we need to first figure out how many individuals will have no children, how many will have one child, how many will have two children, and so forth, when every individual actually has the same chance of having children (i.e., no differential fertility). Recall that when we discussed previously an ideal population and introduced the concept of every individual having the same chance of having children, we likened this to flipping a coin. It turns out that if you flip a coin a certain number of times and want to know the probability of getting a certain number of heads, that probability is given by the **binomial distribution**. Similarly, to figure out how many individuals (out of a population of size N) will have some particular number of children (0, 1, 2, etc.), when everyone has the same probability of having a child, we can use the binomial distribution. The actual formula is not important for our purposes (you can easily look it up on the Internet or in any introductory statistics book if you are interested); what is important is to realize that we can use the binomial distribution to figure this out. Moreover, the binomial distribution can also be used to figure out how N_e is influenced by differential fertility; without going through the gory details, the answer turns out to be:

$$N_e = (4N - 2)/(V_k + 2)$$

where V_k is the variance in the number of children (k) per person.

There's one more thing we need to know before using this formula and that is the average number of children per couple, defined as $\bar{k}$ (whenever you see a symbol with a line on top, you can be pretty sure that it indicates an average value). The above formula actually only holds if the population size is constant. Then what is $\bar{k}$? Hopefully you don't have to think too hard about this to realize that $\bar{k} = 2$ for a constant population size: each couple (consisting of two people) must have on average of two children for the population to neither increase nor decrease. It is not necessary to assume a constant population size to see how differential fertility influences N_e, but it does make the math simpler, and we want to focus on the results, not the math.

Now, to use the formula that tells us how N_e is influenced by differential fertility, let's start by assuming no differential fertility—in other words, that we have a binomial distribution of the number of children per individual. It turns out that the variance for a binomial distribution is $\bar{k}\left(1 - \frac{1}{N}\right)$, so set V_k equal to this in the formula previously and do the algebra (keeping in mind that we are assuming a constant population size, so $\bar{k} = 2$; you should end up with $N_e = N$. No surprise here; since the formula assumes a binomial distribution for the number of children per individual, if in fact we do have a binomial distribution, then there is no differential fertility effect and hence no effect on N_e.

So what happens if we have something other than a binomial distribution? One possibility is a **Poisson distribution**, named after it's discoverer, Siméon Poisson, who came up with it in 1837 while trying to apply probability theory to the decisions of juries in court cases (Poisson must have been a lot of fun at parties, as he was known for frequently saying that life is good for only two things, discovering mathematics and teaching mathematics!). The Poisson distribution is appropriate for events that occur randomly in time or space, such as the number of chocolate chips per chocolate chip cookie, or the length of time spent waiting for a bus to arrive at a bus stop. We will see more examples of the Poisson distribution when we consider the occurrence of DNA mutations over time. In fact, the Poisson distribution is closely related to the binomial distribution, and the variance of a Poisson distribution is equal to the average value, that is, $\bar{k}$. Consider again the variance of a binomial distribution, which is $\bar{k}\left(1 - \frac{1}{N}\right)$: as N gets bigger and bigger, $1/N$ gets smaller and smaller, so the variance of the binomial distribution gets closer and closer to $\bar{k}$, which is the variance of the Poisson distribution. And if we substitute $\bar{k} = 2$ for V_k in the formula for N_e, we get $N_e = N - {}^1/_2$, which is practically the same as N.

Now let's suppose that we have an extremely dictatorial society, in which it is mandated that every couple must have exactly two children—no more, no less. The population size will remain constant, so the average number of children will be $\bar{k} = 2$. What will the variance in the number of children be? A moment's thought should convince you that the variance will be 0—if this is not obvious, go back to the formula for the variance and see what happens when $x_i = \bar{x}$ for every observation x_i. And plugging $V_k = 0$ into the formula for N_e gives the rather astonishing result that $N_e = 2N - 1$; the effective population size is about twice the actual (census) population size! How can this be? Remember that when we defined an ideal population, we said that every individual has the same chance of having children, which is *not* the same as saying that every individual has children. In an ideal population, some people will not have any children, and some will have more than the average number of children. So, for an ideal population, we do expect there to be loss of some alleles in the gene pool over time. But if everyone gets to contribute equally to the gene pool of the next generation, there will actually be more alleles in the gene pool than would be expected for an ideal population of size N. Thus, N_e for such a population is bigger than N.

Let's now consider some real data for differential fertility in human populations. V_k has been estimated for various human populations by counting the number of

children per couple and making some appropriate corrections (e.g., if you count only the number of children per couple at a single point in time, obviously some couples will have more children later, and there are ways to estimate that). For traditional hunter-gatherer societies, V_k is usually around 4, which if you do the math then means that N_e is about $2N/3$. The largest values of V_k that have been observed, for groups with high average fertility such as the Amish, are around 10, for which N_e is about $2N/5$. So, differential fertility in human populations can have an appreciable impact, reducing N by about 30–60%.

N_e FOR HUMANS

Having discussed various factors that can influence N_e, we now ask, what is N_e for humans? We will see in Chapter 5 how one can use genetic data to estimate N_e; for now, it is enough to realize that since the important effect of N_e is to influence the size of the gene pool, the amount of genetic variation in a population is an estimate of the size of the gene pool and can, therefore, (with lots of assumptions!) be used to estimate N_e. This has been done using a variety of methods and data sets for human populations, and the result that is consistently obtained is that N_e for humans is about 10,000. And yet, the census population size for humans is currently about 7 billion, or about five orders of magnitude bigger than N_e. So what is going on here—why is N_e so much smaller than N?

In trying to figure out what might be going on, it is helpful to know whether this large discrepancy between N_e and N is specific for humans, or does it also hold for other species, in particular our closest living relatives, the great apes. It turns out that estimates of N_e for different populations of great apes are about 2–5 times bigger than N_e for humans, even though the census size of great ape populations is much smaller, on the order of a few tens to hundreds of thousands of individuals. If Martian biologists came to Earth, took samples of DNA from humans and the great apes, and estimated N_e, they would conclude that great apes have healthy populations, but they'd better do something about those poor humans, they are clearly in danger of going extinct!

So, compared to great apes, humans are indeed weird in having an N_e that is so many orders of magnitude smaller than N. How might we explain this? If we consider the factors discussed in "The sex ratio and N_e"; "Inbreeding and N_e"; "Variation in population size over time and N_e"; and "Differential fertility and N_e" sections, inbreeding can quickly be ruled out, as levels of inbreeding in human populations are far too low to account for the low N_e. And while there is evidence to indicate that there is both an unequal sex ratio and differential fertility in human populations—in particular, the tendency in many human populations is for fewer males than females to reproduce, with some males having many more children than others—these factors also cannot account for the five orders of magnitude discrepancy between N_e and N. That leaves us with variation in population size over time as the best explanation for the low N_e in humans. In particular, the low N_e indicates that at some point in our past humans must have gone through a **bottleneck** (severe reduction in population size). We will see later on further evidence concerning bottlenecks and their importance in human evolution, but for now it is sobering to realize that the genetic evidence strongly suggests that at some point in our past, there were so few of us that we may have been in danger of going extinct.

CHAPTER 4

A SIMPLE MODEL: HARDY–WEINBERG EQUILIBRIUM

The previous chapter introduced some important concepts about populations, such as ideal populations and the effective population size. The emphasis throughout was on what happens to the gene pool, namely, the kinds of different alleles in a population and their frequencies. We now want to see how various evolutionary forces influence the gene pool. To do so, we will first set up a very simple—and therefore highly unrealistic—model of a population, without any evolutionary forces involved, and see how the gene pool changes over time. In the next chapter, we will make the model more realistic by adding the various evolutionary forces and then see what happens to the gene pool.

Population genetics gets very mathematical very quickly; for those of you who found the math in the last chapter challenging, be forewarned, that was just a taste of what's to come. Still, we can get by with fairly simple algebra, and it is sometimes useful to try to understand the math behind the concepts. Therefore, I will continue to present the equations and urge you to work with them, while also focusing on what we learn from the equations—much of which (hopefully) does make sense, once you think about it.

THE GENE POOL WITH NO EVOLUTION: THE HARDY–WEINBERG PRINCIPLE

Let's start by assuming the following for our population:

- discrete, nonoverlapping generations
- random mating (all matings equally likely to occur)
- infinite population size
- no migration into or out of the population
- no new mutations occur
- everyone has the same **viability** (chance of surviving to the age when reproduction ends) and fertility

This is a pretty boring model! But let's see what happens to the gene pool each generation with such a model. Consider a gene with two alleles, A and B. Let the frequency of $A = p$ and the frequency of $B = q$, then hopefully it is clear that $p + q = 1$. Since humans are diploid, there are three genotypes for the two alleles at this gene: AA, AB, and BB. But in the gene pool model, we don't worry about genotypes; when it comes to producing the next generation, we just consider the alleles. The above assumptions about the population mean that we can think of the production of new offspring as occurring by everyone shedding their gametes (eggs and sperm) into one big pool and then gametes meeting randomly to produce offspring. There are two types of gametes in the pool: A gametes, which have a frequency of p, and B gametes, which have a frequency of q. The chance that two A gametes come together is just $p \times p = p^2$, and this will be the frequency of AA genotypes in the next generation. Similarly, the frequency of BB genotypes corresponds to the chance that two B gametes come together, which is $q \times q = q^2$. The AB heterozygote can come about by either an A gamete from the father and a B gamete from the mother, with frequency $p \times q = pq$, or the other way around, also with frequency pq, so the frequency of AB heterozygotes is $2pq$. And if we add up the frequency of all three genotypes in the next generation, it will be $p^2 + 2pq + q^2$, which can be factored into $(p + q)\ (p + q)$, which

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

equals 1 (since $p + q = 1$), as it should since the sum of the three genotype frequencies must equal 1. Note also that homozygosity thus corresponds to the probability of drawing two alleles that are identical in the gene pool model, while heterozygosity corresponds to the probability of drawing two alleles that are different—we will make further use of this concept in Chapter 10.

And with these genotype frequencies, what will be the allele frequencies in the next generation? To get the frequency of the *A* allele, note that there are two genotypes with this allele: *AA*, which consists entirely of the *A* allele, and *AB*, which is 50% *A* allele. So, to get the frequency of the *A* allele, add up the frequency of *AA* homozygotes (p^2) plus half the frequency of *AB* heterozygotes ($2pq$): $p^2 + pq = p(p + q) = p$. Similarly, the frequency of the *B* allele is obtained by adding the frequency of *BB* homozygotes (q^2) plus half the frequency of *AB* heterozygotes ($2pq$): $q^2 + pq = q(p + q) = q$. So, the frequency of the *A* allele is still p, and the frequency of the *B* allele is still q, which means that in the next generation, the frequency of *AA* homozygotes, *AB* heterozygotes, and *BB* homozygotes will still be p^2, $2pq$, and q^2, respectively. The take-home message: in the absence of any evolutionary forces, allele frequencies and genotype frequencies do not change from generation to generation.

This approach is sometimes referred to as "beanbag genetics," because it is exactly analogous to having a bag with white beans at a frequency p, and black beans at a frequency q, and then asking what is the probability of reaching into the bag twice and getting two white beans, or two black beans, or a white bean and a black bean. For those of you who are skeptical that the beanbag genetics approach really is the same as what actually happens with human reproduction, which involves matings between diploid genotypes, take a look at Box 4.1. Box 4.1 presents the genotype frequencies expected after random mating between diploid genotypes and verifies that with our above population model, randomly drawing pairs of gametes from the gene pool really does have the same outcome as random mating between diploid genotypes

BOX 4.1 ■ Random Mating Involving Diploid Genotypes Is Equivalent to the Random Union of Gametes in the Gene Pool Model

Consider a gene with two alleles *A* and *B* as before, with p designating the frequency of the *A* allele and q designating the frequency of the *B* allele, so $p + q = 1$. Designate the frequency of the *AA* genotype by *D*, the frequency of the *AB* genotype by *H*, and the frequency of the *BB* genotype by *R* (so $D + H + R = 1$). Recall that the frequency of each allele is obtained by taking the frequency of each corresponding homozygous genotype plus half the frequency of the *AB* heterozygotes, so:

$$\text{frequency } (A) = p = D + H/2$$
$$\text{frequency } (B) = q = R + H/2$$

What we then do is consider the frequency of each different type of mating between genotypes and the genotypes that will result in the children of each mating. These are provided in the table below. Look at the first line of the table: matings between *AA* homozygotes will occur with frequency $D \times D = D^2$, and all of the offspring will have the *AA* genotype, so we put D^2 in the *AA* column for the offspring to indicate the overall contribution of this mating to offspring of genotype *AA*. Now look at the second line of the table: matings between an *AA* homozygote and an *AB* heterozygote can occur in two ways (if the male has the *AA* genotype and the female the *AB* genotype, or vice versa), so the overall frequency of such matings is $(D \times H) + (D \times H) = 2DH$. And since half the offspring of this mating will be *AA* homozygotes and half will be *AB* heterozygotes, we put *DH* in the *AA* offspring column and *DH* in the *AB* offspring column. Proceeding in a similar fashion, we fill out the table:

		Offspring genotypes		
Mating	Frequency	*AA*	*AB*	*BB*
$AA \times AA$	D^2	D^2		
$AA \times AB$	$2DH$	DH	DH	
$AA \times BB$	$2DR$		$2DR$	
$AB \times AB$	H^2	$H^2/4$	$H^2/2$	$H^2/4$
$AB \times BB$	$2HR$		HR	HR
$BB \times BB$	R^2			R^2

Note that if we add up the frequencies of all of the different mating types, we get:

$D^2 + H^2 + R^2 + 2DH + 2DR + 2HR = (D + H + R)^2 = 1$, as it should be since we accounted for all of the possible mating types. Now, what do we get in the next generation? Add up the frequencies of each offspring genotype:

$$\text{Frequency } (AA) = D^2 + DH + H^2/4 = (D + H/2)^2 = p^2$$
$$\text{Frequency } (AB) = DH + 2DR + H^2/2 + HR$$
$$= 2(D + H/2)(R + H/2) = 2pq$$
$$\text{Frequency } (BB) = R^2 + HR + H^2/4 = (R + H/2)^2 = q^2$$

So there you have it, random mating of diploid genotypes gives us the same proportions for the genotype frequencies in the next generation as we get under the gene pool model.

(those of you who are happy to accept that the gene pool model is equivalent to random mating between diploid genotypes can skip Box 4.1).

The above facts—namely, that with frequencies of p and q for alleles A and B, the genotype frequencies are p^2 for the AA genotype, $2pq$ for the AB genotype, and q^2 for the BB genotype, and that neither allele nor genotype frequencies change over time—are known as the **Hardy–Weinberg principle**, and the genotype frequencies are sometimes called the Hardy–Weinberg proportions. G. H. Hardy was a British mathematician who was approached by his friend and fellow cricket player, Reginald Punnett (of Punnett square fame, from Chapter 1), for help in figuring out the expected genotype proportions under random mating and all of the other aforementioned assumptions. Hardy quickly worked out the answer, and because he thought this was a trivial result, he didn't try to publish it in the most prestigious scientific journal at that time (*Nature*). Instead, Hardy sent it to a (then) little-known American journal called *Science*, where it was published in 1908 (Hardy 1908). In this note, Hardy made clear what he thought about the mathematical abilities of biologists, stating "… I should have expected the very simple point I wish to make to have been familiar to biologists." Ironically, this "very simple point" has had more lasting influence than any other work by Hardy, and for many years, it was known as Hardy's Principle. In 1943, the geneticist Curt Stern (Stern 1943) pointed out that a German physician by the name of Wilhelm Weinberg had published a more comprehensive derivation (Weinberg 1908) before Hardy published his note, albeit in an obscure German journal (once again, as with Mendel, demonstrating the importance of publishing your work where it will be noticed!). At Stern's suggestion, it has since become known as the Hardy–Weinberg Principle.

Let's work through an example of Hardy–Weinberg to see how it is used. Suppose we have genotype frequencies of 0.25 for AA homozygotes, 0.5 for AB heterozygotes, and 0.25 for BB homozygotes. What are p and q? Recall that to get p (the frequency of the A allele), we take the frequency of AA homozygotes plus half the frequency of AB heterozygotes, which is $0.25 + (0.5/2) = 0.5$. Similarly, to get q, we can take the frequency of BB homozygotes plus half the frequency of AB heterozygotes, which is $0.25 + (0.5/2) = 0.5$. Or, once having obtained p, we can remember that $p + q = 1$, so $q = 1 - p$, which in this example also gives $q = 0.5$ (as it should). And with these allele frequencies, the expected genotype frequencies after random mating (i.e., the Hardy–Weinberg proportions) are:

$$\text{Frequency } (AA) = p^2 = (0.5)^2 = 0.25$$
$$\text{Frequency (AB)} = 2pq = 2(0.5)(0.5) = 0.5$$
$$\text{Frequency (BB)} = q^2 = (0.5)^2 = 0.25$$

So the same genotype frequencies are obtained, and both the genotype and the allele frequencies will stay the same, generation after generation.

I have to confess that when I first learned about the Hardy–Weinberg principle, it seemed circular to me. You start with genotype frequencies, get the allele frequencies, then get the genotype frequencies using the Hardy–Weinberg principle, so it seemed to me that of course these are the same as the genotype frequencies you started with. How could it be any different? For those of you laboring under the same misconception, let's take a few more examples. Suppose we have genotype frequencies of 0.5 for the AA homozygote, 0.5 for the BB homozygote, and no AB heterozygotes at all. What are p and q? Take the frequency of AA homozygotes plus half the frequency of AB heterozygotes, and you get $p = 0.5$, which means that $q = 0.5$, too (do the math if this isn't obvious). And what are the Hardy–Weinberg proportions for these allele frequencies? As we saw before, when $p = q = 0.5$, then the Hardy–Weinberg proportions are 0.25 for the AA homozygotes, 0.5 for the AB heterozygotes, and 0.25 for the BB homozygotes. So we started with only AA and BB homozygotes, but in one generation of random mating, we end up with all three genotypes in their Hardy–Weinberg proportions, with half of the offspring thereby having a genotype (AB) that was completely lacking in the parents.

Let's do one more example. Suppose we now have a population that is 100% AB heterozygotes. This is as different as can be from the above population with 50% AA homozygotes and 50% BB homozygotes, right? And yet, if you do the math, for this population, $p = q = 0.5$, so with one generation of random mating, we get the same Hardy–Weinberg proportions of 0.25 for the AA homozygotes, 0.5 for the AB heterozygotes, and 0.25 for the BB homozygotes. The take-home lesson: there are many different combinations of genotype frequencies that will give the same allele frequencies, but only one combination of genotype frequencies corresponds to the Hardy–Weinberg proportions.

Moreover, no matter how different the genotype frequencies are from the Hardy–Weinberg proportions, all it takes is one generation of random mating (as well as all of the other assumptions listed in the beginning of this chapter) to attain the Hardy–Weinberg proportions. And, once the Hardy–Weinberg proportions are attained, the genotype and allele frequencies remain unchanged, generation after generation, forever and ever. This is, therefore, sometimes referred to as being in Hardy–Weinberg equilibrium. Some of you may be familiar with chemical equilibria, in which the amount of time it takes to get to the equilibrium state depends on how far the existing state of things is from the equilibrium state. Hardy–Weinberg equilibrium is a peculiar kind of equilibrium in that it takes

just one generation to attain the Hardy–Weinberg proportions, no matter how different the current genotype frequencies are from Hardy–Weinberg equilibrium frequencies.

EXCEPTIONS

Having just said that it takes only one generation to attain Hardy–Weinberg equilibrium, I must point out that there are two exceptions to this rule, one trivial, and one more interesting. Both occur when there are different allele frequencies in males and females; the trivial exception involves autosomal genes. To see what happens in this case, consider a simple example in which all the males are *AA* homozygotes and all the females are *BB* homozygotes. Then, with equal numbers of males and females, we have overall a frequency of 0.5 for the *AA* genotype, 0 for the *AB* genotype, and 0.5 for the *BB* genotype. Ordinarily, with these genotype frequencies, we would calculate that $p = q = 0.5$, and, therefore, with one generation of random mating, we should get the Hardy–Weinberg proportions, namely, 0.25 for the *AA* genotype, 0.5 for the *AB* genotype, and 0.25 for the *BB* genotype. But what in fact happens is that all of the matings in the first generation are between male *AA* homozygotes and female *BB* homozygotes; hence, the offspring are 100% *AB* heterozygotes. It is only when these offspring have children that the Hardy–Weinberg proportions will be attained. So, with unequal allele frequencies in males and females for an autosomal gene, it takes two generations of random mating to get to Hardy–Weinberg equilibrium: one generation to get the allele frequencies to be the same in males and females and then one generation to get to Hardy–Weinberg equilibrium.

The more interesting exception to the general rule that it takes only one generation of random mating to get to Hardy–Weinberg equilibrium involves X-linked genes (that is, genes located on the X chromosome). Recall that females are diploid for the X chromosome (having two copies, like the autosomes) but males are haploid, having only one copy of the X chromosome. Since males get their X chromosome from their mother, the frequency of an X-linked allele in males in the next generation is the same as that of females in the present generation. Don't be confused by the fact that males get only one of the two X chromosomes present in their mothers; the overall frequency of an X-linked allele in the X chromosomes transmitted to males is expected to be the same as the frequency of that X-linked allele in the X chromosomes in the mothers. Females get one X chromosome from their mother and one from their father, so the frequency of an X-linked allele in females in the next generation is just the average of the frequencies for that allele in males and females in the present generation. Mathematically, if we let p_M designate the frequency of the X-linked allele in males, and p_F designate the frequency of the X-linked allele in females, we can write:

$$p'_M = p_F$$
$$p'_F = (p_M + p_F)/2$$

where p'_M and p'_F are the allele frequencies in the next generation in males and females, respectively (in general, the "prime" symbol, written "′," denotes the value in the next generation).

If the allele frequencies are the same, then $p_M = p_F$ and then there is nothing of interest, as the allele frequencies in males and females remain the same generation after generation (if this isn't obvious, substitute $p_M = p_F = p$ into the aforementioned equations and verify this for yourself). But suppose the allele frequencies are different in males and females: consider the most extreme case, where $p_M = 1$ and $p_F = 0$ (i.e., males are fixed for an allele that is completely absent in females), shown in Figure 4.1. In the next generation, p_M rather astonishingly plummets from 1 to 0 (because none of the mothers have the allele), while p_F rises from 0 to 0.5. But then in the following generation, p_M rises back up to 0.5, while p_F decreases to 0.25. Over time, p_M and p_F oscillate back and forth, gradually approaching the equilibrium value where $p_M = p_F = 0.33$ (in this case; in general, the equilibrium value will be $\hat{p} = (2p_F + p_M)/3$). So for a sex-linked gene with unequal allele frequencies in males and females, it takes several generations to attain Hardy–Weinberg equilibrium.

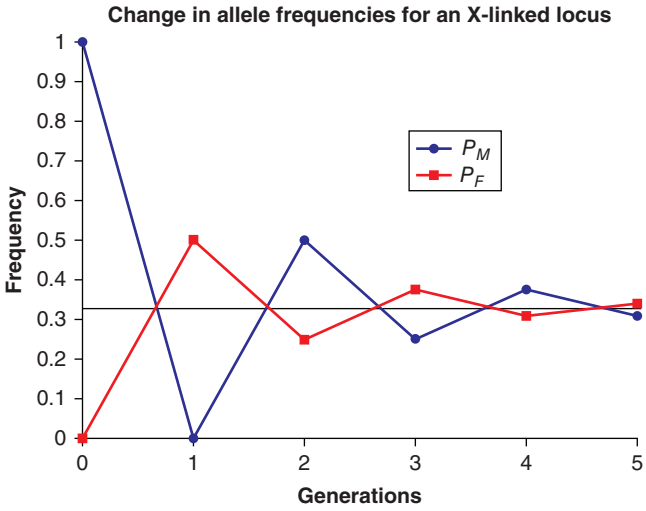


FIGURE 4.1

The change in allele frequencies for an X-linked locus. p_M is the frequency in males and p_F is the frequency in females.

TABLE 4.1 ■ MN blood group data for 16,000 Germans[a]

Genotype	Observed number	Observed frequency	Hardy–Weinberg frequency	Hardy–Weinberg expected number	Chi-square value
MM	4896	0.306	0.304	4866.4	0.180
MN	7856	0.491	0.495	7915.1	0.442
NN	3248	0.203	0.201	3218.4	0.272
Total	16000	1	1	16000	0.894

[a] *Source:* Data taken from Novitski, E., "Genes in Populations," *Human Genetics*, Macmillan Publishing Co., New York, p. 322, 1977.

A REAL-LIFE EXAMPLE

So far we have used contrived examples to illustrate a particular point about the Hardy–Weinberg principle. Let's now see how the Hardy–Weinberg principle is actually used in practice. Consider the following real-life data, from 16,000 Germans typed for the MN blood group. MN is one of the 30 or so additional blood groups (besides ABO and Rh) that humans have, and the MN gene has two alleles (M and N) that are codominant with respect to each other, meaning that all three genotypes (MM, MN, and NN) can be distinguished from one another via the blood group typing. The observed number of each genotype is given in Table 4.1. Are these observed numbers in agreement with those expected assuming Hardy–Weinberg equilibrium? If so, we would conclude that, to a first approximation, the assumptions that underlie the Hardy–Weinberg principle (large population size, random mating, no migration or selection, etc.) probably hold for the *MN* gene in this population. But if the observed numbers for each genotype do not correspond to those expected under Hardy–Weinberg equilibrium, then some additional factor must be influencing the MN genotype frequencies in the German population, and then it becomes interesting to try to figure out what might be causing the discrepancy.

The first thing we have to do is figure out the allele frequencies, for which we need the genotype frequencies. These are easily obtained by dividing the number of each genotype by the total sample size and are given in Table 4.1. We can then compute p (the frequency of the M allele) in the usual way (the frequency of the homozygotes plus half the frequency of the heterozygotes) to get $p = 0.5515$ and, correspondingly, $q = 0.4485$. Actually, we could compute the allele frequencies directly from the observed counts by noting that our sample of 16,000 individuals has 32,000 MN alleles (because this is an autosomal gene), and MM homozygotes have two M alleles while MN heterozygotes have one M allele; therefore, p can be obtained by taking the number of MM homozygotes plus half the number of MN heterozygotes and dividing by 32,000 (i.e., twice the sample size). You can verify for yourself that this gives the same answer; either method works equally well, so it is up to you whether you first calculate genotype frequencies from the observed number of each genotype to get the allele frequencies or calculate allele frequencies directly from the observed number of each genotype.

From the allele frequencies, it is straightforward to calculate the expected Hardy–Weinberg frequency of each genotype (frequency of MM = p^2; frequency of MN = $2pq$, and frequency of NN = q^2) and these are given in Table 4.1. Multiplying these genotype frequencies by the total sample size gives us the expected number of each genotype (assuming Hardy–Weinberg equilibrium), and these are also given in Table 4.1. So now we can answer our question: are the observed genotype numbers the same as those expected assuming Hardy–Weinberg? The answer is well, maybe yes, maybe no: the observed genotype numbers aren't *exactly* the same as the expected genotype numbers, but they are pretty close. And we don't expect the numbers to match exactly, do we, because there will always be chance deviations (flip a coin 10 times and you won't always get five heads and five tails), so now the question is, how close is good enough to say they are the same?

Put another way, how big must the difference between the observed and expected genotype numbers be for us to say that no, they aren't the same? This is a question for statistics, so what we need is a statistical test to tell us how likely it is to get the observed numbers of each genotype when in fact we do have the Hardy–Weinberg proportions. In statistical terms, if this is pretty likely, then we say that we don't have any evidence to reject the hypothesis that we have Hardy–Weinberg proportions. But if the difference is highly unlikely to occur by chance, then we would reject the hypothesis that we have Hardy–Weinberg proportions. It's like flipping a coin 10 times: get six heads and four tails and you'd think you have a fair coin, but if you get 10 heads and no tails, you'd start thinking that you have a two-headed coin.

As it turns out, there are lots of statistical tests that can be used to assess the chance that we would get the observed numbers in Table 4.1 assuming that the Hardy–Weinberg principle holds. Some of these are quite sophisticated, but we'll take a very simple

approach—not because it is the best but because it is good enough to illustrate the underlying idea (which is the most important aspect) while still being simple enough to explain. The particular test we will use is called a **chi-square test** and is usually written with the Greek letter chi as: χ^2. This peculiar name was bestowed by Karl Pearson, one of the founding fathers of statistics, and is based on both the penchant of mathematicians for using Greek symbols and the fact that it has to do with taking the square of numbers (i.e., multiplying a number by itself). To carry out the chi-square test, you take the observed number of each genotype, subtract the expected number, square this number, and then divide the result by the expected number. Do this for each genotype and add up the numbers; the result is the chi-square test value. More formally, the equation is written as follows:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

where o is the observed value, e is the expected value, and the summation is over all observed values.

Inspection of the equation reveals two important features that hopefully make sense. First, the smaller the difference between the observed and expected values, the smaller the chi-square value. So, large values of the chi-square value tell us that the observed and expected values are quite different from one another and, therefore, that the model we used to calculate the expected values is not such a good explanation for our observed data. Second, it is not just the magnitude of the difference between the observed and expected values that is important but rather how large this difference is relative to the expected values. A difference of 10 between the observed and expected values is much more significant if the expected value is 10 than if the expected value is 100.

Table 4.1 includes the calculation of the chi-square value that is obtained when comparing the observed and expected numbers of each genotype. The chi-square value is 0.894; if this number is big enough, we would conclude that the expected numbers do not provide a good fit to the observed numbers, and, therefore, something is wrong with the Hardy–Weinberg model. So, we need to determine how likely we are to get a chi-square value this big, under the assumption that the Hardy–Weinberg model does fit the data. To figure that out, we need to know one more thing, and that is the **degrees of freedom** associated with our chi-square value. Degrees of freedom is usually explained by statisticians as the number of independent classes, or the number of values that are free to vary. However, I find it more convenient to think of the degrees of freedom as the number of observations in the test minus the number of things you need to know in order to calculate the expected values. In our example, we have three observations (the observed number of the three different genotypes), so what do we need to know to calculate the expected values? It turns out that we need to know two things. First, given one allele frequency (p or q), we can obtain the other by subtracting from one, then we can use the Hardy–Weinberg principle to calculate the expected frequency of each genotype. But that is still not enough to carry out the test, because we need to know not just the expected frequency of each genotype but the actual expected number of each genotype, and for that we need to know the sample size. Then we can multiply the expected frequency by the sample size to get the expected number of each genotype and then we can calculate the chi-square value. So, the degrees of freedom associated with our chi-square value is 3 (for the number of genotypes) – 2 (the number of things we need to know, viz., one of the allele frequencies and the sample size) = 1 degree of freedom.

Having our chi-square value of 0.894, and knowing that the associated degrees of freedom are 1, there are two ways we can proceed. The first is to follow the convention that states that if there is a less than 5% chance of obtaining a chi-square value as big (or bigger) than what we obtained, under the assumption that the model used to calculate the expected values is correct, then we conclude that the model is in fact wrong. We can accordingly look up the 5% value (i.e., the value that will be exceeded 5% of the time by chance) in a table of chi-square values with 1 degree of freedom (which you can find on the Internet or in any statistics textbook); this number is 3.84. Since our observed chi-square value is less than this **critical value**, we conclude that our observed values are not significantly different from those expected under the Hardy–Weinberg principle. Thus, we have no evidence to reject the hypothesis that the *MN* blood group gene is in Hardy–Weinberg equilibrium in this population.

This is the way statistical tests are often carried out, but it is important to realize that there is nothing set in stone about the 5% critical value; it is simply a convention that has been adopted. We could just as easily use a different critical value—and indeed, for very important hypotheses, even more stringent critical values may be used, such as 1% or even 0.1%. This is because with a 5% critical value, there is still a 1-in-20 chance that we would get a "significant" chi-square value, even though the model used to obtain the expected values is in fact correct. Moreover, suppose I carry out two tests of Hardy–Weinberg, and in the first case I get a chi-square value of 3.85, while in the second case I get a chi-square value of 3.83. According to the 5% convention, in the first case, I would reject the Hardy–Weinberg model, while in the second, I would not, but these chi-square values are

virtually identical. For these reasons, an alternative approach is to report the probability of the chi-square value (the ***p* value**) if the model is correct, rather than stating whether or not the difference between the observed and expected values is statistically significant. The lower the probability, the worse the fit between the observed and expected values, and it is left up to you to decide whether the p value is low enough to reject the model used to calculate the expected values. For the data in Table 4.1, the p value is 0.64, meaning that there is about a 64% chance of getting differences between the observed values and the expected values as big (or bigger) than the observed differences, when the Hardy–Weinberg model is in fact correct. This is certainly high enough that we would judge that we have a good fit between the Hardy–Weinberg proportions and our observed values. Since this is more informative than simply stating that we either reject or do not reject our model on the basis of a 5% critical value, this is the approach we will often use in this book.

One final comment about statistics before we move on and that is to keep in mind that statistics is a numbers game. Flip a coin 10 times and get six heads and four tails, and you would conclude that you have a fair coin. Flip a coin 1000 times and get 600 heads and 400 tails, and—in addition to having a sore thumb—you would conclude that you do not have a fair coin (do the chi-square test, if you like!). Yet, in both cases, the proportion of heads is the same, namely, 60%. That is why, strictly speaking, we do not say that our statistical test leads us to accept the model that we used to calculate the expected values for the test. Instead, we say that on the basis of the data we obtained, we do not have enough evidence to reject the model we used to calculate the expected values. Maybe the model is true, or maybe if we had more data (in particular, a bigger sample size—e.g., more flips of the coin), we would reject the model. That is why it is always important to keep in mind the **power** of a statistical test—that is, how big a deviation from the expected values is needed in order to have enough evidence to reject the model? It is often the case that with the sample sizes typically obtained in molecular anthropological studies, the statistical tests have very little power—that is, only very large deviations from the model can be detected. Smaller, subtler deviations cannot be detected. But as the next section shows, this apparent drawback can sometimes be a blessing in disguise.

SOME PRACTICAL USES FOR HARDY–WEINBERG

Even though the assumptions underlying the Hardy–Weinberg principle (random mating, large population size, no migration, no selection, etc.) seem highly unrealistic, especially when applied to humans, in the vast majority of cases, one finds a good fit between observed genotype frequencies and those expected at Hardy–Weinberg equilibrium. This is partly because of the power issue of the statistical tests discussed in the previous paragraph and partly because the departures from the underlying assumptions of the Hardy–Weinberg model that occur in humans often turn out not to have such a big impact on the observed genotype frequencies. Moreover, when the observed genotype frequencies do *not* match those expected at Hardy–Weinberg equilibrium, the reason almost always turns out to be mistakes in the genotyping. In fact, mistakes in genotyping account for such a large portion of observed deviations from Hardy–Weinberg proportions that many studies will automatically assume that any deviations must be due to such mistakes. Thus, to a first approximation, the usual expectation is that observed genotype proportions will match those expected under Hardy–Weinberg equilibrium, and this expectation can then be used to make further inferences.

For example, suppose we have data for a recessive trait (where the phenotype of the heterozygote cannot be distinguished from one of the homozygotes), such as the rhesus (Rh) blood group, and we want to estimate the frequencies of the Rh+ and Rh− alleles. Let's use a concrete example: suppose we perform Rh blood typing on a group of people and we find that 84% are Rh+ and 16% are Rh−. What are the frequencies of the Rh+ and Rh− alleles? Recall that to get the allele frequency, you take the frequency of the homozygotes plus half the frequency of the heterozygotes. But all we know from the blood typing is that 84% are Rh+, some of which are Rh+/Rh+ homozygotes and some of which are Rh+/Rh− heterozygotes. With no way of knowing which are which, we can't figure out the allele frequencies. But if we can assume Hardy–Weinberg equilibrium, then it is easy: the frequency of the recessive phenotype (in this case the Rh− blood type) is expected to be the frequency of homozygotes for the recessive allele. So, if q is the frequency of the Rh− allele, then the frequency of the Rh− blood type is q^2, and in this case, we get $q = \sqrt{0.16} = 0.4$, and $p = 1 - q = 0.6$. In fact, this simple method—which assumes Hardy–Weinberg equilibrium—is commonly used to estimate allele frequencies for recessive traits.

An interesting historical example that made good use of the Hardy–Weinberg principle is something that we took for granted in the first chapter, namely, figuring out the inheritance of the ABO blood groups. Actually, after the discovery of the ABO blood groups at the beginning of the twentieth century, there was some uncertainty as to how they were inherited. The majority view was that the blood groups were controlled by two genes, each with two alleles. According to this view, there was an A gene with two alleles (A

TABLE 4.2 ■ ABO blood group genotypes under the 2-locus and 1-locus models[a]

Blood type	Genotypes: 2-locus model	Genotypes: 1-locus model
A	A-bb	AA, AO
B	aaB-	BB, BO
AB	A-B-	AB
O	aabb	OO

[a]The "-" indicates that either of the two alleles at that locus can be present.

TABLE 4.3 ■ ABO blood types observed in an African Pygmy population and the expected numbers under the 2-locus and 1-locus models[a]

Blood type	Observed	Expected 2-locus model	Expected 1-locus model
O	88	93.2	89.3
A	44	38.9	42.8
B	27	21.8	25.7
AB	4	9.1	5.2
Total	163	163	163

[a]*Source:* From Bernstein, F., "Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstructuren des Menschen," *Klinische Wochenschrift* 3:1495, 1924.

and a, with A dominant over a), and a *B* gene with two alleles (B and b, with B dominant over b). The ABO blood type that would result from each possible genotype is shown in Table 4.2; the basic idea is that if you have at least one dominant A allele but are homozygous recessive at the *B* gene, you will have blood type A; if you are homozygous recessive at the *A* gene but have at least one B allele, you will have blood type B. Blood type AB would result from having at least one A allele and at least one B allele, while individuals who were homozygous recessive at both the *A* and *B* genes (i.e., aabb) would have blood type O. This scheme is entirely sensible, especially when you remember that this was before anyone knew what a gene actually was; the prevailing view was that genes by their very nature could have only two alleles, corresponding to "on" and "off." So, the dominant allele would represent the "on" state, and the recessive allele would represent the "off" state.

The minority view was that the ABO blood groups were controlled by a single gene with three alleles, corresponding to A, B, and O. We've already covered how these three alleles determine the four ABO blood types, in Chapter 1, but for convenience, this information is repeated in Table 4.2. The controversy as to which view was correct continued until 1924, when the German mathematician Felix Bernstein carried out an elegant statistical analysis to determine which hypothesis best explained the observed data (Bernstein 1924, 1925). In order to do so, he assumed that Hardy–Weinberg equilibrium held, which enabled him to come up with estimates of the allele frequencies for both models. Bernstein then took real data, computed the expected frequencies of the four blood groups under both models, and then compared the expected values to the observed values to determine which model best fits the data. The actual computations involve a lot of algebra that is not terribly illuminating, so we'll skip the gory details and go right to the results. Table 4.3 gives some data on the ABO blood groups in an African Pygmy group, along with the expected values computed under either the 2-locus, 2-allele model or the 1-locus, 3-allele model. Just by looking at the values, you can see that the 1-locus model is a much closer fit than the 2-locus model, and in fact the chi-square test gives a much lower *p* value for the 2-locus model than for the 1-locus model. So, based on Bernstein's elegant use of the Hardy–Weinberg principle, the familiar 1-locus model of inheritance was accepted for the ABO blood groups.

Now, those of you who have been paying attention should be thinking, what a load of nonsense, you don't need any "elegant" statistical analysis that assumes Hardy–Weinberg equilibrium to figure this out, because there is a straightforward difference between these two models in the predicted offspring genotypes from a particular kind of family. Have you spotted it? Consider what should happen with families where one parent is blood type AB and the other is blood type O. Under the 2-locus model, the AB parent can have one of four possible genotypes (AABB, AaBB, AABb, AaBb) while the O parent can have only one genotype (aabb). Depending on the genotype of the AB parent, there are four potential kinds of gametes that this parent can produce: AB, Ab, aB, and ab, which when combined with the ab gamete from the O parent would result in blood types in the children of AB, A, B, and O, respectively. Although we don't know what to expect in any given family, the point is that if we study lots and lots of families with one AB parent and one O parent, we should expect to see all four blood groups among the offspring if the 2-locus model is correct.

Now consider what we would expect in such families under the 1-locus model. The AB parent is genotype AB and, therefore, produces A and B gametes, while the O parent is genotype OO and produces only O gametes. We, therefore, expect the offspring of such families to be about 50% blood type A and 50% blood type B, and that's it—we shouldn't see any AB or O children. So, to figure out whether the 2-locus or 1-locus model of inheritance is correct, you can skip all the complicated algebra and "sophisticated" statistical

TABLE 4.4 ■ Blood types observed among the offspring of families where one parent is type AB and the other is type O, before and after Bernstein's study was published in 1924[a]

	O	A	B	AB
Before 1924	27	80	59	24
After 1924	2	228	234	1

[a] *Source:* From Sturtevant, A.H., *A History of Genetics*, Harper & Row: New York, NY, 1965.

analysis and just look at the blood groups in the children of lots of families where one parent is type AB and the other is type O. If you see children with all four blood types, then the 2-locus model is correct; if you see only blood types A and B, then the 1-locus model is correct. Pretty simple, right?

So why didn't the geneticists at the time figure this out, instead of having to rely on a mathematician to set them straight? The answer is that they did indeed figure this out—but they got the wrong answer! Table 4.4 shows a compilation of the observed blood types in the offspring of AB and O parents for the years both before and after Bernstein published his analysis. There is a dramatic difference: before Bernstein published his analysis, there were many offspring with blood groups AB or O, but after Bernstein's analysis came out, hardly any offspring with blood groups AB or O were observed. There were no significant technical improvements in blood typing methodology during this time that might explain this discrepancy. Instead, it appears that this is a classic example of theory driving data: before Bernstein, the experimentalists "knew" that the 2-locus theory was correct (remember, it was the majority view); therefore, they knew that they had better seen some AB and O offspring in these families. While we don't know for sure how they managed this, there are always a few individuals who give indeterminate results for the blood typing—maybe their RBCs react a bit more or a bit less with an antibody than most people do—and one might thus be tempted to classify these individuals as having blood groups that one expects to see. But after Bernstein, the experimentalists "knew" that the 1-locus theory was correct, and hence they should not see any AB or O offspring. And so, they didn't.

Except that if you look at Table 4.4, in fact there were still a few AB or O offspring observed after Bernstein published his analysis. How can we account for these exceptions? The first explanation that should always come to mind whenever an unusual result is observed is that it most likely is a mistake. But let's suppose we repeat the typing on these exceptional individuals and get the same result. What else might explain these unexpected AB and O blood types in the offspring? Another possible explanation is that maybe they reflect a new mutation. While this is possible, it turns out that mutations occur at such a low rate (remember this from Chapter 2) that the number of unexpected AB and O blood types is too high to be explained by new mutations.

In fact, the most likely explanation for these unexpected AB and O blood types among the children is **nonpaternity**: the actual father of such a child is not the supposed father, and hence has a different genotype, which accounts for the unexpected blood group of the child. Estimates of the amount of nonpaternity in humans are not easy to come by (for obvious reasons!), and they can vary quite widely, but a reasonable guesstimate is about 5% nonpaternity in your average urban society. Now, 5% may not sound like very much, but to put this number in perspective, when I used to teach an introductory course in biological anthropology at Penn State University, there would usually be about 300 students enrolled, and 5% of 300 means that roughly 15 students in the course had a different daddy than they think they did. Nonpaternity is something that every human geneticist who works with family data has to be aware of as a potential explanation for exceptional results. It is also the reason why most high school biology/genetics courses no longer carry out any blood typing or other typing of genetic polymorphisms among the students (as we did in my biology class when I was a high school student), because sooner or later somebody is going to get an unpleasant surprise (which, in fact, is precisely why blood typing of students in the biology course at my high school was later discontinued!).

This finishes our discussion of the Hardy–Weinberg principle, which remains an extremely useful (albeit very simple) model of how genes are expected to behave in populations. We also introduced some important concepts, such as statistical tests of a model or hypothesis, that we will make extensive use of later. In the next chapter, we will see what happens when we relax the assumptions underlying the Hardy–Weinberg principle.

CHAPTER 5

EVOLUTIONARY FORCES

In the previous chapter we developed a very simple—albeit unrealistic—model of the gene pool in the absence of any evolutionary forces and saw that this model leads to Hardy–Weinberg equilibrium. Now, we want to see what happens to the gene pool when we allow various evolutionary forces—that is, factors that can influence the composition of the gene pool—to occur. In particular, we are interested in the following questions:

1. What happens to the genetic variation within a population when we violate the assumptions of the Hardy–Weinberg model: does it increase, decrease, or stay the same?
2. What happens to the genetic differences among populations: do they increase, decrease, or stay the same?

And why are we interested in these questions? Because, as we shall see later, the genetic variation within populations and the genetic differences among populations provide key insights into the genetic history of populations, which is one of the major goals of molecular anthropology. Thus, to properly interpret patterns of genetic variation and genetic differences, we need to understand how these are influenced by the various evolutionary forces.

So, let's begin by recalling what happens to the gene pool under the Hardy–Weinberg model. Although the examples in the previous chapter were based on a gene with two alleles, it is easy to extend the Hardy–Weinberg proportions to any number of alleles. For a gene with n alleles, we can designate the alleles as $A_1, A_2, A_3, \ldots, A_n$, each with frequency $p_1, p_2, \ldots, p_n$, respectively (and, therefore, $p_1 + p_2 + \ldots + p_n = 1$). Given a set of genotype frequencies, we can get p_1 by taking the frequency of A_1A_1 homozygotes plus half the frequency of all heterozygous genotypes that include the A_1 allele. At Hardy–Weinberg equilibrium, the expected frequency of each genotype is:

$$\text{Homozygotes for allele } A_i : \text{frequency}(A_iA_i) = p_i^2$$
$$\text{Heterozygotes for alleles } A_i \text{ and } A_j \ (i \neq j) : \text{frequency}(A_iA_j) = 2p_ip_j$$

And once these equilibrium frequencies are attained (which usually takes just one generation), they stay the same forever and ever—as long as the assumptions used to derive the Hardy–Weinberg proportions hold. Recall that these assumptions are:

- Discrete, nonoverlapping generations.
- Random mating (all matings equally likely to occur).
- Infinite population size.
- No new mutations occur.
- No migration into or out of the population.
- Everyone has the same viability and fertility.

So, what happens when we relax these assumptions? The first one is easy; with continuous generations instead of discrete generations, everything still works out to be the same. It's just that the math becomes a lot harder—with continuous generations, you need calculus and differential equations to figure things out, whereas with discrete generations, we can get by with algebra. And since we're not interested in the mathematical details, we'll take the easy way out and stick with discrete generations in what follows.

NON–RANDOM MATING

Instead of everyone having the same chance of mating, what happens with non–random mating? There are two kinds of non–random mating to consider: **assortative mating** and **inbreeding**. Assortative mating simply means preferential mating and can occur either as **positive assortative mating**, meaning that

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

matings between people who are alike in some aspect of their phenotype occur more frequently than expected by chance, or as **negative assortative mating**, meaning that matings between people who differ in some aspect of their phenotype occur more frequently than expected by chance. Negative assortative mating is very rare in humans, with one obvious exception—see whether you can come up with it. I'll provide the answer at the end of this section.

Positive assortative mating, on the contrary, is quite common in humans—so much more so than negative assortative mating, that we usually just use "assortative mating" to refer to positive assortative mating. The way assortative mating is measured is by calculating the **correlation coefficient** between husbands and wives for a particular trait. Correlation coefficients can vary between −1 and 1, where positive values indicate that husbands and wives are more alike for a trait than expected, negative values indicate that husbands and wives are more different than expected, and values near zero indicate no correlation. Some typical correlation coefficients between husbands and wives are around +0.3 for height, +0.4 for political preference, and +0.5 for education (Schwartz 2013). It should come as no surprise that we tend to choose as mates individuals with backgrounds, beliefs, and so forth, that are similar to our own. That is not to say that all couples are completely alike in every respect—we all know couples who are so completely at odds over everything that you wonder how they ever got together—just that there is a significant tendency for couples to have similar backgrounds and beliefs.

How does assortative mating influence the gene pool? For most genetic variation, the answer is probably not at all, because we don't consider a person's genotype when it comes to choosing a mate. After all, nobody makes such a decision based on a prospective mate's ABO blood group genotype! But if there is a genetic component underlying a phenotype for which there is assortative mating (such as height), then there can be an effect on the gene pool. It turns out that this effect is basically the same as what happens with the other type of non–random mating, namely, inbreeding, so we'll discuss that now. But first, the answer to the question posed previously: the one obvious example of negative assortative mating in humans is sex, as by definition all matings in humans involve people of different sex (at least, all matings that lead to offspring, which is all we are concerned with in this book).

Inbreeding, also known as **consanguinity**, involves mating between individuals who are related (as introduced back in Chapter 1). Inbreeding can thus be thought of as matings between individuals with genotypes that are more similar than expected under random mating. And since assortative mating occurs between individuals with similar phenotypes, to the extent that genotypes influence those phenotypes, then the expected consequences of assortative mating and inbreeding for the gene pool are the same. It turns out that inbreeding is often associated with reduced viability and fertility. Why is this the case? Let's see what happens to the gene pool with inbreeding.

We start by defining F, the **inbreeding coefficient**, as the probability that the two alleles at a locus in an individual are **identical by descent**, meaning that they are both descended from a common ancestor in a previous generation (if this definition sounds familiar, we also used this definition in Chapter 3). Note that while this means that the genotype in such cases will be homozygous for the allele in question, this is not the only way homozygotes can occur. Homozygotes can also occur if the alleles are the same just by chance, not because they come from the same ancestor, and this is called **identity by state**. For a concrete example, refer to Figure 5.1.

Some of you may already be familiar with the principle that all of the alleles at a locus trace back to a single common ancestor—which we will discuss in more detail in Chapter 12—and hence may find the distinction between identity by descent versus identity by state confusing. If all of the variation at a locus traces back to a single common ancestor at some point in the past, then aren't all alleles by definition identical by descent? The answer is that when considering inbreeding, we are interested only in people who are related in the past few generations. Anything involving more distant relationships is not considered inbreeding, because the consequences for the gene pool are negligible, and, therefore, any identical alleles coming from such distant relatives are by definition identical by state.

So how does inbreeding influence the gene pool? In Box 5.1, I go through how to derive the genotype frequencies with inbreeding; those of you who don't care about such details can just proceed directly to the equations, which are:

$$\text{Frequency (AA)} = p^2 + pqF$$
$$\text{Frequency (AB)} = 2pq(1 - F)$$
$$\text{Frequency (BB)} = q^2 + pqF$$

Note that when there is no inbreeding, $F = 0$, and the aforementioned equations reduce to the familiar Hardy–Weinberg proportions. But when there is inbreeding, $F > 0$, we add pqF to the frequency of each homozygote and subtract $2pqF$ from the frequency of the heterozygotes. And since p, q, and F are all positive numbers, this means that homozygosity is increased relative to the Hardy–Weinberg proportions, while heterozygosity is decreased, by a factor of $1 - F$. So another interpretation of the inbreeding coefficient is that it measures the proportion by which the

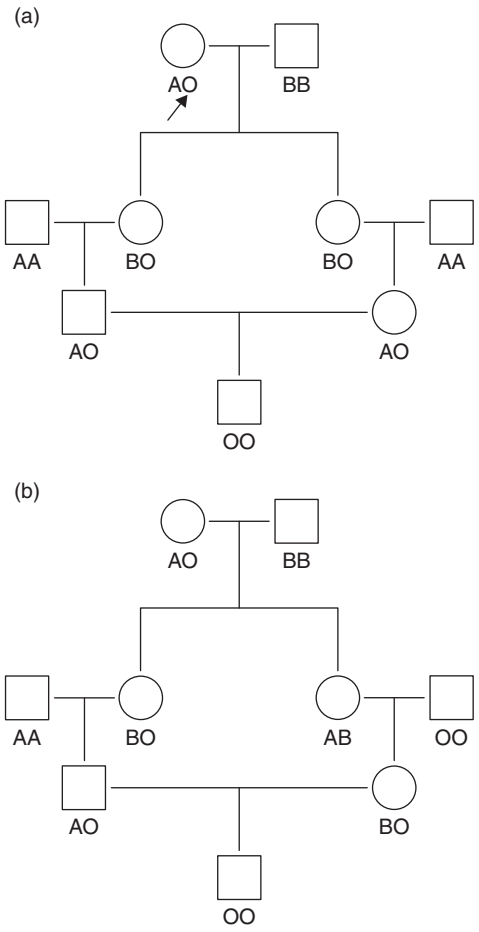


FIGURE 5.1

Pedigree of an individual whose parents are first cousins, showing the difference between identity by descent and identity by state. (a) Identity by descent: suppose this individual has ABO blood type O and is, therefore, homozygous for the O allele. One way this could happen would be if an O allele in one of the great grandparents (arrow) is transmitted to both grandparents, and each grandparent transmits the O allele to the parents of the individual, who then transmit the O allele in turn to the individual. (b) Identity by state: the individual is still homozygous for the O allele, but in this case, the O alleles came from unrelated individuals.

heterozygosity is reduced relative to a randomly mating population with identical allele frequencies.

And what happens to the allele frequencies after one generation of inbreeding? Recall that the frequency of the A allele will be equal to the frequency of AA homozygotes plus half the frequency of AB heterozygotes:

$$\begin{aligned} p' &= p^2 + pqF + 0.5\ (2pq\ (1 - F)) \\ &= p^2 + pqF + pq - pqF \\ &= p^2 + pq = p(p + q) = p \end{aligned}$$

So, the allele frequencies do not change with inbreeding, just the genotype frequencies. And the genotype frequencies change such that the frequency of homozygotes is increased while the frequency of heterozygotes is decreased. In the most extreme case, when $F = 1$ (which corresponds to self-mating, not something that we have to worry about with humans, but does occur in some organisms such as some plants or nematodes), the population will consist only of homozygotes, with no heterozygotes. The most common types of matings between relatives in humans involve uncle–niece, first cousins, or second cousins, with inbreeding coefficients of 0.125, 0.0625, and 0.0156, respectively. Incidentally, for those of you who—like myself—get confused by the distinction between first cousins once removed and second cousins, the easy way to remember this is that first cousins have grandparents in common, second cousins have great grandparents in common, third cousins have great great grandparents in common, and so forth, whereas children of your first cousin are your first cousin once removed, children of your first cousin's children are your first cousin twice removed, and so forth.

With these inbreeding coefficients, there are generally only modest increases in homozygosity (I leave this for you to work out for yourself). Still, this is enough to explain why inbreeding causes decreased viability and fertility, namely, increased homozygosity for deleterious (harmful) recessive alleles. It turns out that each of us is heterozygous for a few dozen or so mutations that would have a negative impact on health and/or fertility if they were to become homozygous. So, anything that increases homozygosity is likely to increase the frequency of deleterious traits.

Just how deleterious are the effects of inbreeding in humans? This was a subject of intense personal interest for Charles Darwin, because he married his first cousin, Emma, and had 10 children with her—of which three died in childhood. Darwin lobbied unsuccessfully to have questions about cousin marriage and possible consequences included on the 1871 Census of Great Britain. It was left to his son, George Darwin, to carry out the first systematic study of the effects of inbreeding. George found that the frequency of individuals with parents who were first cousins was about 3.5% among all students at Cambridge and Oxford but only 2.4% among students who participated in the sport of rowing (Darwin 1875). So there you have it: if your parents are first cousins, then you are less likely to be fit enough to participate in rowing!

There have been numerous subsequent studies of the effect of inbreeding on other aspects of human health, viability, and fertility. Some of the most detailed information came inadvertently from U.S.-sponsored studies of the effects of radiation on the survivors of

BOX 5.1 ■ Expected Genotype Frequencies with Inbreeding

Start by assuming two alleles at a locus, A and B, with frequencies p and q, respectively. Then under our gene pool model, we draw two alleles at random to make offspring. What are the chances of an AA homozygote? The first allele drawn must be an A, which happens with probability p. If the second allele is also an A, it could be identical to the first A allele by descent, which happens with probability F. Or, it could be identical in state, which happens if the allele is both not identical by descent (with probability $1 - F$) and an A allele (with probability p). Thus,

Frequency (AA)
$= pF + p^2(1 - F)$ multiply to get rid of the parentheses
$= pF + p^2 - p^2F$ rearrange
$= p^2 + pF - p^2F$ factor pF from the last two terms
$= p^2 + pF(1 - p)$ recall $1 - p = q$
$= p^2 + pqF$

Similar reasoning holds for the frequency of BB homozygotes, so

$$\text{Frequency (BB)} = q^2 + pqF$$

Finally, to get the frequency of AB heterozygotes, it is simplest to recall that

Frequency (AB)
$= 1 - (\text{frequency (AA)} + \text{frequency (BB)})$
$= 1 - (p^2 + pqF + q^2 + pqF)$ recall $p^2 + 2pq + q^2 = 1$
$= p^2 + 2pq + q^2 - p^2 - pqF - q^2 - pqF$ do the arithmetic
$= 2pq - 2pqF$ factor out $2pq$
$= 2pq(1 - F)$

Q.E.D.

the atomic bombs dropped on Japan. These studies did not find any convincing evidence for an increase in mutations among children of exposed individuals compared to unexposed controls (Schull et al. 1981). However, the principal investigators, James Neel and William Schull, realized that there was a relatively high frequency of first cousin marriages (about 7%) in the Japanese population as a whole, which enabled an in-depth investigation of the effects of inbreeding. They documented rather modest increases in the frequency of stillbirths and childhood mortality when the parents were related; for example, the overall death rate for children was around 10–11% when the parents were first cousins, versus 8–9% for unrelated parents (Neel and Schull 1962).

However, inbreeding can be quite common in some communities; for example, among some Bedouin and Asian Indian communities, the frequency of marriages involving relatives can reach 20–50% (Romeo and Bittles 2014). And there have been some provocative claims that such high levels of inbreeding over multiple generations can lead to "purging" of deleterious recessive alleles due to the increased homozygosity associated with inbreeding, thereby exposing such alleles to selection (as discussed in the "Selection" section) and reducing their frequency. However, such claims remain controversial, as there are many other sociodemographic factors that often differ between marriages involving relatives versus marriages involving unrelated people that can have an impact on childhood mortality—marriages among relatives is a common and convenient way of keeping wealth in the family, for example, so in many societies, the children of related parents tend to be born into better circumstances than the average child.

Finally, what is the effect of inbreeding on genetic variation within populations and genetic differences between populations? You may at first think that nothing happens to genetic variation, because inbreeding in and of itself does not result in the loss of alleles, so you might think that no loss of alleles means no loss of genetic variation. But remember, what inbreeding does is increase homozygosity and decrease heterozygosity, and we equate variation with heterozygosity, so, therefore, the result of inbreeding is that genetic variation (aka heterozygosity) is indeed decreased within populations. And since the effects of inbreeding are random, the heterozygosity lost in one population because of inbreeding will be different from the heterozygosity that is lost in another population. Thus, the genetic differences between populations will increase.

SMALL POPULATION SIZE

The next departure from the assumptions of the Hardy–Weinberg model to consider involves population size. The Hardy–Weinberg model assumes an infinitely large population size; what happens if instead we have a small population size? There are three important consequences: (1) random fluctuations in allele frequencies become important; (2) there is loss of genetic variation; and (3) inbreeding becomes more important. Let's consider each of these in turn.

Small Populations: Random Fluctuations in Allele Frequencies

Recall that with Hardy–Weinberg, allele and genotype frequencies both remain constant over time. However,

if the population size is small enough, then the allele frequency can change from generation to generation, just because of chance effects in which alleles are sampled from the gene pool. These random fluctuations in allele frequencies are known as **genetic drift**. It is easy enough to model the effects of small population size on the sampling of alleles. The low-tech way is to get a bag and two different colors of beans or beads, say black and white, and put in 10 of each color. This mimics a population of 10 diploid individuals with two alleles at a locus, each with a frequency of 0.5. Draw 20 beads from the bag, with replacement (so put each bead back into the bag after you draw it), to mimic the production of 10 offspring in this population (so, we keep the population size constant), and keep track of how many black and white beads you get. Suppose you get 12 black beads and 8 white beads; now empty out the bag, put in 12 black and 8 white beads and repeat the process until you get bored or until you end up with all black or all white beads.

The hi-tech way to model genetic drift is to use a computer simulation. There are several good ones available for free on the Internet; one I can recommend as I use it when teaching is called allele A1, available from the following Web site: http://faculty.washington.edu/herronjc/SoftwareFolder/AlleleA1.html.

In addition to the program, a manual and some useful tutorials can be downloaded as well. Some sample outcomes of genetic drift are provided in Figure 5.2. Look over Figure 5.2, taking note of the starting allele frequency and population size for each set of graphs. It should be evident that the amount of fluctuation in the allele frequency is related to population size. The bigger the population, the smaller the fluctuation. So, the smaller the population, the bigger the impact of genetic drift in terms of changing allele frequencies.

Small Population Size: Loss of Genetic Variation

With genetic drift, alleles always end up becoming either lost (frequency = 0) or fixed (frequency = 1). This may not be completely evident, as in some graphs in Figure 5.2 both alleles are still present at the end of the simulation. But the take-home point is that if you run the simulations long enough, you will always end up with loss or fixation of an allele. So, genetic drift results in loss of genetic variation over time. The expected rate at which this happens can be determined mathematically; the details are provided in Box 5.2 for those who are interested. The equation that describes how heterozygosity is lost over time is:

$$H_t = (1 - 1/2\mathrm{N})^t H_0$$

where H_t is the heterozygosity in the current generation, H_0 is the heterozygosity in some arbitrary initial generation (e.g., when a small number of people leave a large population to start a new population), N is the (effective) population size, and t is the number of generations. The smaller N is, the bigger $1/2N$ is, and hence the smaller $(1 - 1/2N)$ is. Figure 5.3 shows graphs of the loss of heterozygosity over time for several values of N. The overall message: the smaller the population, the bigger—and faster—the loss of heterozygosity.

Another important point is that the probability that an allele will be fixed rather than lost is related to the initial frequency of the allele. Alleles that start at high frequency have a correspondingly higher probability of becoming fixed rather than lost. In fact, it turns out that the probability that an allele will ultimately become fixed is just the frequency of the allele. An allele with a frequency of 0.5 has a 50:50 chance of becoming fixed. An allele with a frequency of 0.9 has a 90% chance of becoming fixed. So what does this tell you about the chance that a newly arising mutation will become fixed in the population by genetic drift? Well, it depends on the effective size of the population. A newly arising mutation is, by definition, present in just one copy. So if the population has effective size $N = 50$, then a new mutation has a frequency of 1%, and thus there is a 1% chance that this new mutation will drift to fixation. And if the population has $N = 500$, then there is just a 0.1% chance that this new mutation will drift to fixation. The vast majority of the time, newly arising mutations simply get lost by genetic drift.

So in terms of genetic variation within populations, and genetic differences between populations, what does genetic drift do? Alleles become lost by genetic drift, so hopefully it is clear that genetic variation within populations decreases because of genetic drift. And since different alleles are lost and fixed by genetic drift in different populations (because the process is random), genetic drift increases the genetic differences between populations.

Small Population Size: Increase in Inbreeding

The third consequence of small population size mentioned in the "Small population size" section is that it can lead to an increase in inbreeding, because in a small population many individuals will be related. Thus, even "random" matings are likely to involve relatives, and therefore we expect an increase in the proportion of alleles that are identical by descent (i.e., the inbreeding coefficient) over time in a small population. The details as to how this happens are provided in Box 5.2 for those who are interested; the relevant equation is:

$$F_t = 1 - (1 - 1/2N)^t$$

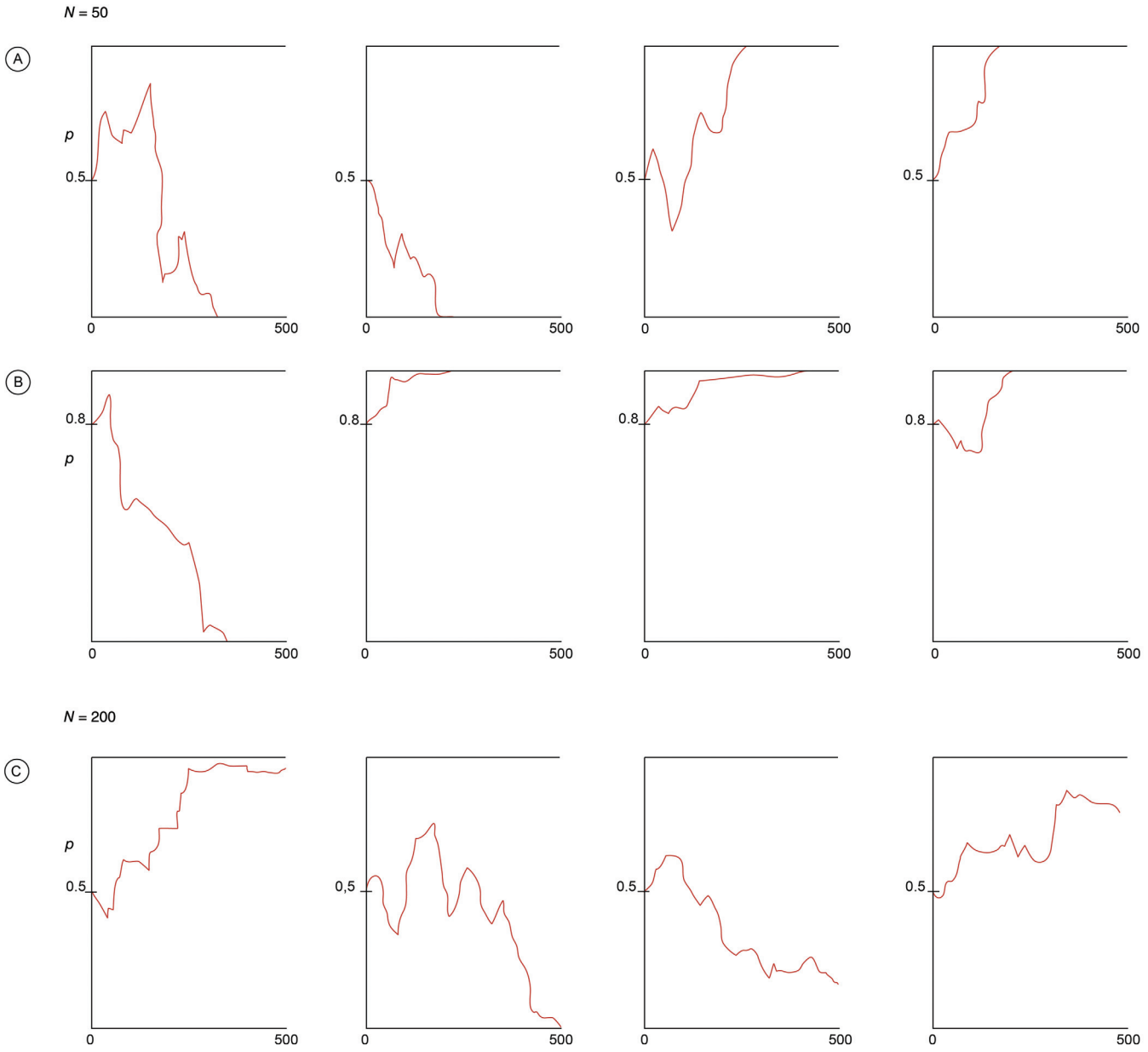


FIGURE 5.2

Genetic drift, which is the random fluctuation in allele frequencies over time. Each row shows four outcomes from computer simulations for a particular combination of population size (N) and allele frequency (p). A: $N = 50$, $p = 0.5$. B: $N = 50$, $p = 0.8$. C: $N = 200$, $p = 0.5$.

where F_t is the inbreeding coefficient (i.e., the probability that two alleles drawn from the gene pool at random are identical by descent) in the present generation, assuming that there was no inbreeding initially (so $F_0 = 0$) and that the population has been small (of effective size N) for t generations. Figure 5.4 illustrates how inbreeding increases over time for various population sizes; as with loss of heterozygosity, the smaller the population, the bigger—and faster—the increase in inbreeding.

Example of Small Population Size: Tristan da Cunha

The classic, textbook example of a small human population is Tristan da Cunha, a tiny island (less than 100 km^2) located in the middle of nowhere (actually, in the south Atlantic some 2800 km from the nearest land, South Africa). Named after the first European to set eyes on the island, the Portuguese explorer Tristão da Cunha, Tristan da Cunha was taken over by the British in 1816, who stationed a garrison of

BOX 5.2 ■ Loss of Heterozygosity and Increase in Inbreeding in Small Populations

Start with our standard gene pool model of two alleles (A, B) at a locus with frequencies p and q, respectively. Only now we also have to take into account the size of the gene pool. Suppose we have a population with effective size N; we, therefore, have $2N$ alleles in our gene pool. Recall that F is the probability that two alleles are identical by descent, so let F_t be this probability in the current generation t. To produce offspring, we draw two alleles from the gene pool. The probability that the second allele is identical to the first allele is $1/2N$. The probability that the second allele is different from the first allele is thus $(1 - 1/2N)$, but these two alleles may be identical by descent in the previous generation, with probability F_{t-1}. Therefore,

$F_t = 1/2N + (1 - 1/2N)F_{t-1}$ multiply both sides by -1 and add 1

$1 - F_t = 1 - 1/2N - (1 - 1/2N)F_{t-1}$ factor $(1 - 1/2N)$ from the right side

$1 - F_t = (1 - 1/2N)(1 - F_{t-1})$

Recall that with inbreeding, heterozygosity $= 2pq(1 - F)$. Let H_t be the heterozygosity in generation t, then substituting into the above equation (note that the $2pq$ term cancels from both sides):

$$H_t = (1 - 1/2N)H_{t-1}$$

Note that using the same logic, $H_{t-1} = (1 - 1/2N)\ H_{t-2}$, substitute this into the equation:

$$H_t = (1 - 1/2N)^2 H_{t-2}$$

And $H_{t-2} = (1 - 1/2N)H_{t-3}$, so we can keep doing this. This is an example of a recurrence equation, in which the value of a term in a sequence is a function of the previous terms. Here, the heterozygosity in the present generation can be written as a function of the heterozygosity in the previous generation, which can in turn be written as a function of the heterozygosity in the generation before that, and so on. If we assume some arbitrary starting generation at time $t = 0$ with heterozygosity H_0, we get:

$$H_t = (1 - 1/2N)^t H_0$$

Q.E.D.

To see what happens to inbreeding in a small population, go back to the equation we derived for the inbreeding coefficient in the current generation, F_t:

$$1 - F_t = (1 - 1/2N)(1 - F_{t-1})$$

Note that this is also a recurrence equation, so we can obtain the equation for F_t in terms of the inbreeding in the "initial" generation. The details I leave as an exercise for you; you should get:

$$1 - F_t = (1 - 1/2N)^t(1 - F_0)$$

And if we assume that there was no inbreeding in the initial generation—which would be the case if, for example, our small population was founded by a few individuals coming from a much larger population without any inbreeding—then $F_0 = 0$, and we end up with:

$$F_t = 1 - (1 - 1/2N)^t$$

Q.E.D.

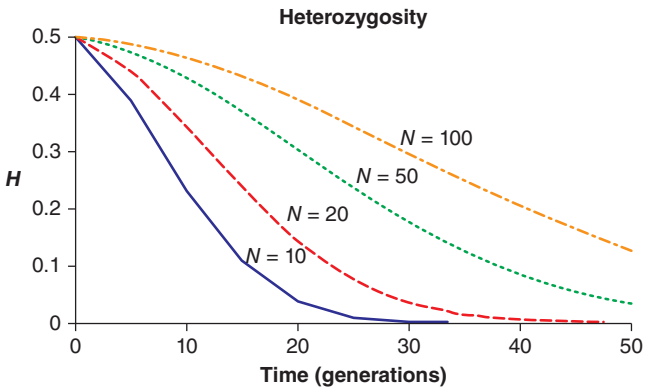


FIGURE 5.3

Loss of heterozygosity (H) over time as a function of population size (N).

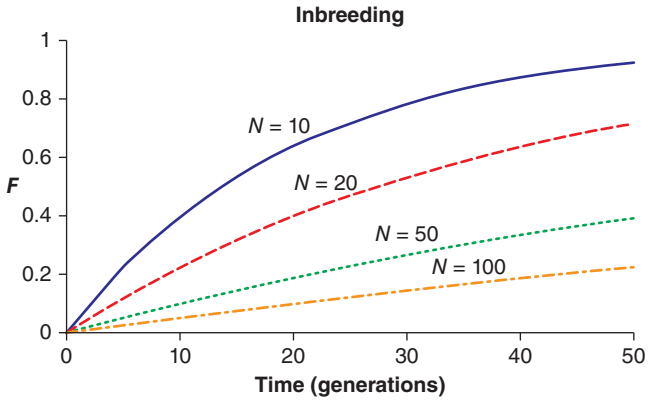


FIGURE 5.4

Increase in inbreeding (F) over time as a function of population size (N).

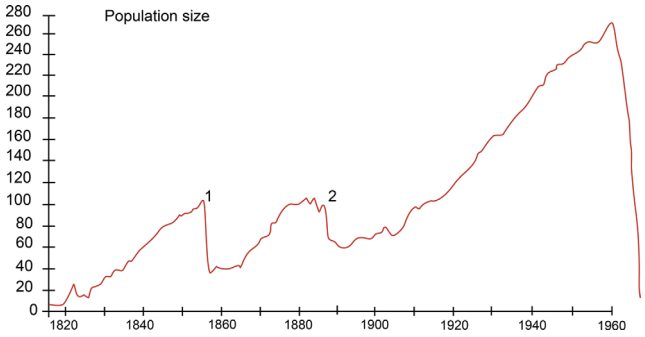


FIGURE 5.5

Population size change over time for Tristan da Cunha. The numbers 1 and 2 indicate sudden sharp decreases in population size, as explained in the text. Modified with permission from Roberts, D.F., "Genetic effects of population size reduction," *Nature* 220:1084, 1968.

soldiers there as a deterrent to any attempt to rescue Napoleon Bonaparte from his exile on St. Helena (some 2000 km away!). It took only 2 years for the British military to decide that having soldiers stationed 2000 km from St. Helena probably wasn't much of a deterrent, so in 1818 the garrison was removed. However, some soldiers requested and were granted permission to stay on the island, along with their wives (mostly prostitutes from Cape Town), and this was the beginning of a permanent settlement. From church and other records, the entire history of the island has been worked out. Figure 5.5 shows the population size each year and shows that overall there has been consistent growth in the population, with some notable population declines. Most of the population growth reflects births on the island from the starting population of about 15 people, with the occasional addition of shipwrecked sailors or others who came to the island voluntarily (before the age of steamships, Tristan da Cunha was an important watering place for passing ships).

The first period of population reduction reflects two events that happened in 1856–1857. First, one of the original founders died, who apparently exerted a strong will when he was alive, as after his death some 25 of his descendants left the island. Second, the first missionary to arrive on the island (in 1851) evidently was not persuaded by the charms of living there, as he left a few years later with another 45 people, stating that "it will be a happy day when this little lonely spot is once more left to those who probably always were … its only fit inhabitants—the wild birds of the ocean" (Roberts 1968).

But the population recovered from these events and continued growing, until disaster struck in 1885. The island lacks a natural harbor, so the islanders had to row out in small boats to meet ships that called. And in 1885, a boat with 15 adult males was on the way to meet a vessel when a sudden storm arose, and the boat sank with the loss of all 15 men. This left a grand total of four adult men on the island, including one who was mentally ill and two who were elderly! Tristan became an island of widows, and during the next few years many women left with their children; by 1891 the population had decreased by nearly half.

But even this disaster was not enough to kill off the population, as the population resumed growing—until October 1961, when without any warning a volcano thought to be dormant began erupting on the island. The entire population was evacuated to Britain, where they remained for about a year, until the volcano ceased erupting and the island was declared safe for people to return. The islanders were given the choice to either stay in Britain or return to Tristan, and how many do you think chose to return? Every single one! Which, I guess, tells you something about what it is like to live in Britain…. Today about 260 people live on Tristan, making their living from farming, a lobster fishery, and the sale of commemorative coins and stamps to collectors.

The detailed records for Tristan not only make it possible to reconstruct the history of population size changes but also the increase in inbreeding. Inbreeding has increased at a regular rate, to the point where nearly everyone in the population is inbred, and the average inbreeding coefficient is about 0.05—meaning that everyone is nearly as closely related as first cousins. It is likely that the inbreeding coefficient would be even higher, if not for the few shipwrecked sailors and others who came to the island during the early years, thereby contributing unrelated alleles to the gene pool. And this increase in inbreeding arose not because people chose to mate with relatives but rather—as can be documented from the records—because among the available partners, all were related. Curiously, though, no negative health effects have been documented that can be specifically related to inbreeding, and even homozygosity at blood group loci is not as high as one might expect.

Although in some respects Tristan is a special case, in others it probably is representative of many early human populations. A small band of people leave their group and colonize a new place, facing many dangers that threaten the extinction of the population. Although many such groups undoubtedly did go extinct, those that survived built up their population size to sustainable levels, albeit with an increase in inbreeding. The potential negative effects of inbreeding would have been ameliorated by occasional migration into the population (as will be explored in more detail later in this chapter), bringing in new alleles. Such was likely the situation for most human populations—at least, before the domestication of plants and animals,

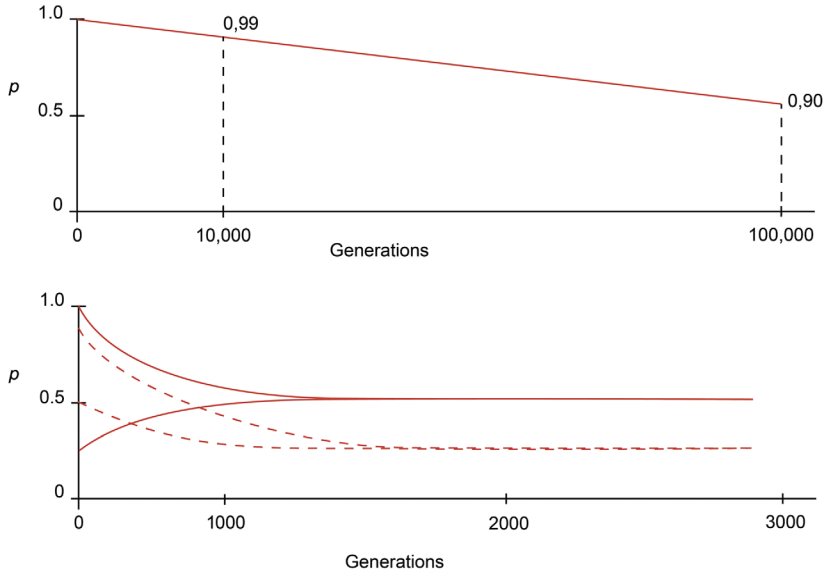


FIGURE 5.6

Change in allele frequency (p) over time due to mutation. Top diagram: allele frequency change with only forward mutation (i.e., loss of the allele due to mutation) and a starting frequency of $p = 1.0$. The graph is based on the following equation: $p_t = (1 - \mu)^t p_0$ where p_0 is the initial frequency of allele A, p_t is the frequency of allele A after t generations, and μ is the mutation rate. For those of you who are interested in how this equation was derived, I leave that as an exercise for you. Hint: to get an A allele in the present generation, you have to draw an A allele from the gene pool that has *not* mutated. Bottom diagram: allele frequency change with both forward and reverse mutation for various starting values of p. Solid lines, with both mutation rates = 1×10^{-3}; dashed lines, with forward rate = 1×10^{-3} and back mutation rate = 1×10^{-4}.

which then fueled much larger human expansions, as discussed later in this book.

MUTATION

The next assumption of Hardy–Weinberg equilibrium that we will investigate is mutation. What happens when we allow new mutations to occur? On the one hand, mutation is the most important evolutionary force, because mutation alone is responsible for creating new variation. And without variation, there can be no evolution. Without variation, none of the other evolutionary forces (non–random mating, genetic drift, migration, or selection) have anything to act on. But paradoxically, as we shall see below, while mutation reigns supreme among evolutionary forces in terms of creating variation, when it comes to changing allele frequencies, mutation is the least important evolutionary force. This is because mutation rates typically are very low.

Let's see what happens in our gene pool model when we allow mutations to occur. Start as usual with two alleles, A and B, with frequencies p and q, respectively. Suppose we start with $p = 1$ (i.e., the population is fixed for the A allele), and we allow A alleles to mutate to B alleles at a rate of 1×10^{-6} per generation—that is, one out of every million A alleles mutates to a B allele (which is a reasonably high mutation rate). With just mutation in one direction (from A to B, but not from B to A), what do you expect to happen? Hopefully a moment's thought is enough to convince you that eventually every A allele will mutate to a B allele, and hence the population will become fixed for the B allele. And how long will this take? Figure 5.6 shows the change in the frequency of the A allele over time; after 10,000 generations (about 250,000–300,000 years for humans, with a generation time of 25–30 years), the frequency of the A allele is still 0.99—a change of just 1% in roughly the amount of time our species, *Homo sapiens*, has existed! And after 100,000 generations, the frequency of the A allele is still 0.9. It will take some 4.7 *million* generations of one-way mutation—over 100 million years, about the amount of time that mammals have been around!—for the frequency of the A allele to decrease to 0.01. The take-home lesson: because mutation rates are so low, they are pretty much useless when it comes to changing allele frequencies.

A more realistic model would be to allow mutation both from A alleles to B alleles (forward mutation) and from B alleles to A alleles (back mutation). Figure 5.6

shows what happens when the mutation rates are the same in both directions. Regardless of the starting frequency of the A allele, the same equilibrium value of 0.5 is reached. So, the equilibrium value does not depend on the allele frequencies but only on the forward and back mutation rates. In Figure 5.6, unrealistically high mutation rates of 1×10^{-3} are used in order to show that equilibrium is reached during a reasonable time span (still, a few thousand generations!). With more realistic, lower mutation rates, it will take even longer to reach equilibrium. And if the forward and back mutation rates differ, then an equilibrium value that depends only on the mutation rates will still be reached that will not be 0.5 but some other value (Figure 5.6).

To see why this should be so, let's go back to our gene pool model. Now we have mutation from A alleles to B alleles occurring at a rate μ, and back mutation from B alleles to A alleles at a rate ν. In order to get an A allele from our gene pool in the present generation, we have to draw either an A allele from the gene pool that does not mutate (with probability $p(1 - \mu)$) or a B allele that does mutate (with probability $q\ \nu$). Therefore, the frequency of the A allele in generation t is:

$$p_t = p_{t-1}(1 - \mu) + q_{t-1}\nu$$

where the $t - 1$ subscript refers to the allele frequencies in the previous generation. So then what happens over time is that eventually the gain of new A alleles via mutation from B alleles is exactly balanced by the loss of A alleles via mutation to B alleles. When this happens, $p_t = p_{t-1} = \hat{p}$, the equilibrium allele frequency. Substitute this into the above equation, keep in mind that $\hat{q} = 1 - \hat{p}$, and do the arithmetic; you should get:

$$\hat{p} = \frac{\nu}{\mu + \nu}$$

As we saw in Figure 5.6, the equilibrium allele frequencies depend only on the forward and backward mutation rates, not on the initial allele frequencies. However, it will still take a very very *very* long time to ever reach equilibrium, again because mutation rates are so low.

Equilibrium between Genetic Drift and Mutation

Recall that genetic drift (random changes in allele frequencies from generation to generation) ultimately leads to homozygosity and loss of genetic variation, whereas mutation will keep introducing new alleles. What happens when we consider both genetic drift and mutation? At some point, we should reach an equilibrium where the loss of variation due to genetic drift is exactly balanced by the gain of variation due to new mutations. Note that in the previous section, we considered just two alleles at a locus and assumed that mutations change one of these two alleles into the other. In the rest of this section, we will take a more realistic view of mutations: each new mutation within a locus causes a completely new allele to exist (this is called the **infinite alleles** model). While this assumption is violated by certain kinds of mutations, in practice it holds pretty well. This is because most genetic loci consist of thousands to hundreds of thousands of nucleotide positions, so the chance of getting the same mutation at the same nucleotide position within a locus is pretty small. Box 5.3 goes through the math to derive the equation for the heterozygosity at equilibrium under the infinite alleles model, so take a look if you are interested in the details. Otherwise, here is the equation:

$$\widehat{H} = \frac{4N\mu}{4N\mu + 1}$$

where $\widehat{H}$ is the heterozygosity at equilibrium, N is the effective population size, and μ is the mutation rate. Figure 5.7 plots $\widehat{H}$ versus $4N\mu$, and hopefully the relationship shown in the figure makes sense. As either N or μ gets bigger, then $\widehat{H}$ also gets bigger: this happens because less variation is lost via genetic drift in larger populations, while more variation is introduced with a higher mutation rate. Conversely, with smaller values of $4N\mu$ then $\widehat{H}$ also is smaller, because more variation is lost via genetic drift in smaller populations, while less variation is introduced with a lower mutation rate. In fact, once $4N\mu$ is above 10 or below 1, then either practically everyone is heterozygous, or practically nobody is heterozygous. Given the importance of $4N\mu$ in determining how much variation there is in a population, it has been given its own symbol (a Greek letter, of course!): Θ (theta).

Since the above equation for heterozygosity with genetic drift and mutation does not assume any natural selection (discussed in a later section in this Chapter), this is sometimes referred to as **neutral evolution** or **neutrality**, the idea being that genetic variation in this model is "neutral" with respect to an individual's chances of surviving and reproducing. This expected relationship has therefore sometimes been used to test whether the variation observed at a particular gene (or set of genes) is consistent with the predictions of neutrality. The idea is that if we know—or have an estimate of—N and μ, then we can estimate Θ and, in turn, get an estimate of $\widehat{H}$, which we can then compare to our observed estimate of heterozygosity. Sounds good—but the problem with this approach is that N (the effective population size) is a large number with a lot of uncertainty (it could be thousands to hundreds of thousands or more), and μ is a small number with a lot of uncertainty (1 in a 10 million to 1 in a billion,

BOX 5.3 ■ Equilibrium between Genetic Drift and Mutation

We start with the equation we derived in Box 5.2 for the probability that two alleles are identical by descent in a small population:

$$F_t = 1/2N + (1 - 1/2N)F_{t-1}$$

Under the infinite alleles model, every new mutation introduces an allele that does not already exist in the population. Therefore, in order to draw two alleles that are identical by descent from the gene pool, neither allele has mutated, so we add a factor of $(1 - \mu)^2$ to the above equation:

$$F_t = (1/2N + (1 - 1/2N)F_{t-1})(1 - \mu)^2$$

At equilibrium, the increase in homozygosity due to the small population size (measured by F) is, by definition, exactly balanced by the decrease in homozygosity due to new mutations, so $F_t = F_{t-1} = \hat{F}$. Substitute this into the equation and multiply everything out:

$$\begin{aligned}\hat{F} &= (1/2N + \hat{F} - \hat{F}/2N)(1 - 2\mu + \mu^2) \quad \text{keep multiplying} \\ &= 1/2N + \hat{F} - \hat{F}/2N - \mu/N - 2\hat{F}\mu + \hat{F}\mu/N + \mu^2/2N \\ &\quad + \hat{F}\mu^2 - \hat{F}\mu^2/2N\end{aligned}$$

This looks like a mess, but let's see what we can do. First, remember that μ is typically a very small number, on the order of 10^{-6} to 10^{-8} or so, so μ^2 is practically 0, which means that we can ignore terms with μ^2. That gets us:

$$\hat{F} = 1/2N + \hat{F} - \hat{F}/2N - \mu/N - 2\hat{F}\mu + \hat{F}\mu/N$$

Now subtract $\hat{F}$ from both sides and get everything over the least common denominator, namely, $2N$:

$$\begin{aligned}0 &= (1 - \hat{F} - 2\mu - 4N\mu\hat{F} + 2\hat{F}\mu)/2N \quad \text{multiply both sides by } 2N \text{ and rearrange} \\ &= 1 - 2\mu - \hat{F} - 4N\mu\hat{F} + 2\hat{F}\mu\end{aligned}$$

Note that $1 - 2\mu \approx 1$ for typical small values of μ:

$$= 1 - \hat{F} - 4N\mu\hat{F} + 2\hat{F}\mu$$

Take the terms involving $\hat{F}$ over to the other side and factor out $\hat{F}$:

$$\hat{F}(1 + 4N\mu - 2\mu) = 1 \quad \text{set } 1 - 2\mu \text{ again equal to 1 and divide both sides by } (1 + 4N\mu)$$

$$\hat{F} = \frac{1}{1 + 4N\mu}$$

We can define the heterozygosity ($\hat{H}$) at equilibrium to be $1 - \hat{F}$, so $\hat{F} = 1 - \hat{H}$:

$$1 - \hat{H} = \frac{1}{1 + 4N\mu}$$

and

$$\hat{H} = \frac{4N\mu}{4N\mu + 1}$$

Q.E.D.

more or less), and when you multiply a large uncertain number by a small uncertain number, you end up with a very uncertain number. What tends to work better in practice is to take estimates of $\hat{H}$ (which we can measure fairly precisely) and μ (which we are getting better at measuring) and use these to estimate N. In fact, this is a common way of estimating the effective population size, although there is still inevitably a lot of uncertainty surrounding such estimates.

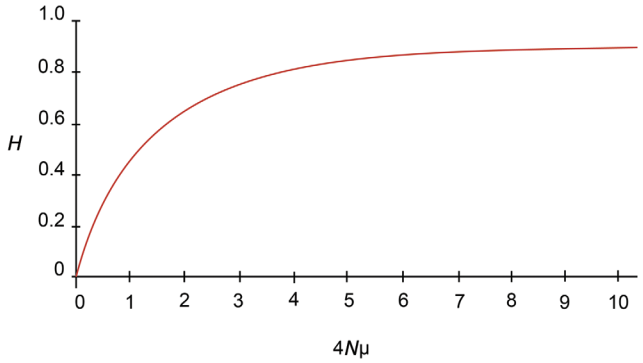


FIGURE 5.7

Graph of heterozygosity (H) versus $4N\mu$, where N is the effective population size and μ is the mutation rate.

Rate of Neutral Evolution

We've just seen what happens to overall heterozygosity in a population as alleles are lost by genetic drift and generated by new mutations. We can also ask what happens to any particular new mutation. The vast majority of the time, a new mutation will be lost via genetic drift in the first few generations, but occasionally a new mutation will rise in frequency. And once in a very great while, a new mutation will reach fixation, completely replacing the allele(s) that were originally present. The fixation of a new mutation—purely by genetic drift, without any natural selection involved—is referred to as the **rate of neutral evolution**. How often does this occur?

Suppose we have a population with effective size N, and neutral mutation rate μ (by neutral mutation rate,

we mean the rate of new mutations that are neither advantageous nor disadvantageous to have). Then,

Number of alleles in the population = $2N$

Number of new mutations per generation = $2N\mu$

Probability that a new mutation will reach fixation = $1/2N$ (recall that the probability of fixation of an allele via genetic drift is equal to the frequency of the allele, and by definition a new mutation is present in just one copy in the gene pool).

Rate of fixation of new mutations = (rate at which new mutations arise) (probability of fixation) = $(2N\mu)(1/2N) = \mu$.

So we end up with the surprising result that the rate of neutral evolution is simply equal to the neutral mutation rate. We'll discuss neutral evolution (and other aspects of molecular evolution) later on in Chapter 6, but there is one further point to make here: note that the rate does not depend at all on population size, which may seem counterintuitive. The reason is that in a small population, there are fewer new mutations occurring each generation, but fixation goes more quickly. Conversely, in a big population, there are more new mutations occurring each generation but fixation takes longer. Remarkably, these two processes balance each other exactly, so the overall rate of neutral evolution is the same regardless of the population size.

Figure 5.8 shows the overall process of neutral evolution in a population. New mutations arise ($2N\mu$ of them every generation), of which the vast majority are lost via genetic drift. But on average every $1/\mu$ generations a new mutation will keep rising in frequency and reach fixation. How long does it take a mutation that is going to be fixed to reach fixation? The derivation of this result is too complex to go into, so I will just tell you the answer: it takes on average $4N$ generations for the fixation to occur, which hopefully makes sense, as the larger the population, the longer it takes for fixation to occur.

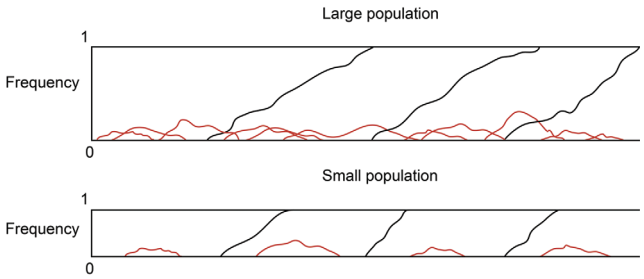


FIGURE 5.8

Fixation of neutral mutations. In a large population (top), many mutations occur but only a few rise to fixation. In a small population (bottom), fewer mutations occur, but more of them rise to fixation. Overall, the rate of fixation of neutral mutations is independent of population size.

Finally, what is the effect of mutation on genetic variation within populations, and genetic differences between populations? Hopefully, this is an easy one for you. Mutation increases genetic variation within populations, as new alleles are continually being generated. And since the mutations that occur in one population are different from those in another population, mutation increases the genetic differences between populations. But keep in mind that when it comes to changing existing allele frequencies in a population, mutation takes a back seat to all of the other evolutionary forces discussed in this chapter.

MIGRATION

The next departure from the assumptions of Hardy–Weinberg equilibrium to discuss is migration (also known as **gene flow**), the movement of individuals (and thus, alleles) between different populations. We will start with a simple model of one-way migration shown in Figure 5.9. The source population has alleles A and B with frequencies P and Q, respectively, while the recipient population has the same two alleles with frequencies p and q. Every generation, a fraction m of the alleles in the recipient population is contributed by the source population and the remaining 1-m fraction of the alleles comes from the recipient population. What do you suppose will happen under this model? Hopefully, a moment's thought will convince you that with one-way migration, eventually p and q will converge to P and Q, respectively. The equation that describes how this happens is (with the derivation in Box 5.4, for those who are interested):

$$p_t - P = (1 - m)^t(p_0 - P)$$

where p_t is the allele frequency in generation t, and p_0 is the allele frequency when migration begins. The assumption here is that the allele frequencies in the

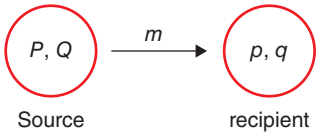


FIGURE 5.9

A simple model of one-way migration. Each generation, a source population with allele frequencies P and Q (so $P + Q = 1$), contributes a fraction m (where m is between 0 and 1) migrants to the recipient population with allele frequencies p and q (so $p + q = 1$).

BOX 5.4 ■ One-Way Migration (Admixture)

Start with a source population with alleles A and B at frequencies P and Q, respectively. We assume that P and Q do not change over time. The recipient population has these same two alleles at frequencies of p and q, respectively, and receives migrants from the source population at a rate of m per generation. With our gene pool model, there are two ways of getting an A allele in the recipient population in the next generation: either the A allele is a migrant from the source population, with probability mP, or it was present in the recipient population in the previous generation, with probability $(1 - m)p_{t-1}$. Thus,

$$p_t = (1 - m)p_{t-1} + mP \quad \text{subtract } P \text{ from both sides}$$
$$p_t - P = (1 - m)p_{t-1} + mP - P \quad \text{factor } (1 - m) \text{ on the right side}$$
$$p_t - P = (1 - m)(p_{t-1} - P)$$

Note that $p_{t-1} - P = (1 - m)(p_{t-2} - P)$, and so on, leading to the recurrence relationship:

$$p_t - P = (1 - m)^t(p_0 - P)$$

where p_0 is the allele frequency in the recipient population when the admixture starts.

Extending this equation to the exchange of migrants among multiple populations is straightforward. Consider the island model depicted in Figure 5.11. We assume that each population is large enough that genetic drift can be ignored; our usual two alleles (A and B) with average frequencies across all populations of P and Q; and that each population exchanges migrants with all other populations at a rate m. Then m is simply the probability that an allele in one population will be from a migrant. As with one-way migration, the probability of an A allele in generation t in any population is the probability that the allele came from the same population in the previous generation, which is $(1 - m)p_{t-1}$, plus the probability that the allele came from a different population, which is mP, leading to:

$$p_t = (1 - m)p_{t-1} + mP$$

Solve this as for one-way migration to get:

$$p_t - P = (1 - m)^t(p_0 - P)$$

Q.E.D.

source population (P and Q) do not change over time. The left side of the equation ($p_t - P$) is the difference in the allele frequency in generation t between the recipient population and the source population. The right side of the equation has a similar expression ($p_0 - P$), which is the initial difference in the allele frequency between the recipient and source populations. If there is no migration, then the initial difference in the allele frequencies never changes, which hopefully makes sense: we want to know what changes when there is migration, so with no migration, no change. And what does happen when there is migration? Each generation, the difference in allele frequencies gets reduced by a factor of $(1 - m)$. And since m is a positive number between 0 and 1, over time the factor $(1 - m)^t$ approaches 0, and so the difference in allele frequencies ($p_t - P$) also approaches 0.

How fast does this happen? Figure 5.10 plots the change in the allele frequency in the recipient population for various values of the migration rate. The lesson here is that, in contrast to mutation, migration changes allele frequencies very quickly. Even for what would be considered low rates of migration, it takes only a few hundred generations for the allele frequency in the recipient population to become identical to that in the source population. And the reason why migration is so much faster than mutation at changing allele frequencies is because migration rates are typically much higher than mutation rates. A very low migration rate would be on the order of 0.001 (i.e., 1 out of every 1000 alleles in the recipient population are contributed by the source population each generation), while a very high mutation rate would still be a few orders of magnitude lower than this.

How realistic is this one-way model of migration? Actually, there are many situations involving humans where migration occurs mostly—if not exclusively—in one direction. This sort of migration is often referred to as **admixture**; a classic example is the case of African–Americans. Ancestors of African–Americans were brought forcibly to the United States to serve as slaves, and during the slavery period, there was predominantly gene flow from European–Americans

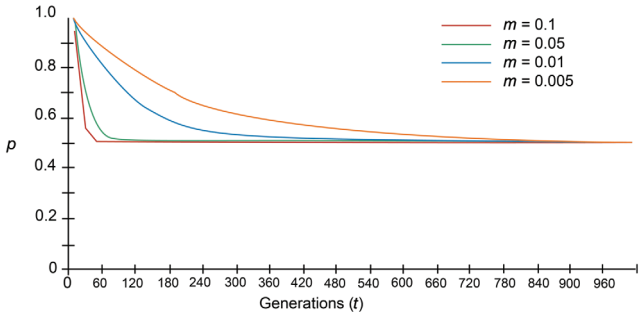


FIGURE 5.10

Change in allele frequency (p) in the recipient population over time with various migration rates (m). At the start, p in the recipient population is 1.0 and P in the source population is 0.5.

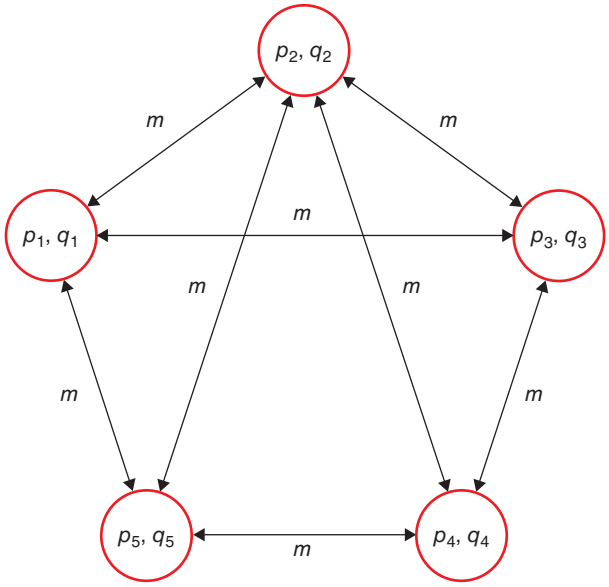


FIGURE 5.11

A simple island model of migration among subpopulations. p_i and q_i are the frequencies of the A and B alleles in subpopulation i. Over time, p_i and q_i will converge on the average allele frequencies over all subpopulations, P and Q.

into the African–American gene pool, with very little gene flow going in the other direction. We can use genetic data and our model of one-way migration to estimate how much gene flow occurred. In doing so, we assume that the allele frequency in European–Americans today is the same as it was in the past, and similarly that the allele frequency in West African populations today is the same as in the Africans brought to the United States. Whether these assumptions are actually true or not we don't know, but most scientists won't let an untestable assumption or two stand in the way of an analysis that they want to do, and neither will we.

The most useful alleles for estimating admixture are those that show large differences between the source and recipient populations—if the allele frequencies are the same to start within the source and recipient populations, then estimating admixture is rather hopeless! A particularly useful allele for our purpose here is the Duffy blood group allele Fy^a. Named after a hemophiliac patient who responded to multiple blood transfusions by making a novel antibody, anti-Fy^a, the Duffy blood group system includes several antigens and plays an important role in resistance to the *Plasmodium vivax* malaria parasite. One study (Workman et al. 1963) found that the frequency of the Fy^a allele is 0.422 in European–Americans, 0.0 in West Africans, and 0.045 in African–Americans. Assuming that gene flow from European–Americans into African–Americans began 10 generations ago, what is the estimated migration rate per generation of European–American alleles into the African–American gene pool? From this information, we have $P = 0.422$, $p_0 = 0.0$, $p_t = 0.045$, and $t = 10$. Substituting these values into the equation gives us:

$$0.045 - 0.422 = (1 - m)^{10}\,(0.0 - 0.422)$$

which reduces to

$$0.893 = (1 - m)^{10}$$

For those of you unfamiliar with equations of this form, there are two ways to solve this for m. The first is the time-honored method called "guess and check"; start by guessing some value of m and then see how close $(1 - m)^{10}$ is to 0.893 and then keep refining your guess until you get the answer you want. Don't laugh—guess and check can be a lot faster and easier than trying to solve some complicated equation, especially if you are a good guesser. Otherwise, you can solve this with logarithms. The logarithm of any number is equal to the exponent that we have to raise 10 to in order to get that number. For example, $\log(100) = 2$, because $10^2 = 100$. In actuality, logarithms can be to any base, but for our purposes base 10 works just fine, so we'll stick to that. To solve the equation, start by taking the log of both sides:

$\log(0.893) = \log(1 - m)^{10}$ in general, $\log(x)^a = a\log(x)$
$-0.049 = 10\log(1 - m)$ divide both sides by 10
$-0.0049 = \log(1 - m)$ take the antilog of both sides (i.e., raise 10 to the power of both sides; the antilog of $\log(x) = x$)
$0.989 = 1 - m$ and so
$m = 0.011$

Therefore, we estimate that the rate of admixture from European–Americans into the African–American gene pool has been about 1.1% per generation.

Extending this model of one-way migration to migration among several populations is straightforward. The underlying model is depicted in Figure 5.11, where we assume several populations, each of which is large enough that we can ignore genetic drift. We further assume that each population exchanges genes with all other populations at a rate m, and that we have two alleles, A and B, with average frequencies across all populations of P and Q, respectively. This model is sometimes referred to as an **island model** of population structure, and with this model the relevant equation is (with details in Box 5.4):

$$p_t - P = (1 - m)^t(p_0 - P)$$

This is similar to the previous equation for one-way migration, but now p refers to the allele frequency in

BOX 5.5 ■ Migration Versus Genetic Drift

We start with the equation we derived in Box 5.2 for the probability that two alleles are identical by descent in a small population:

$$F_t = 1/2N + (1 - 1/2N)F_{t-1}$$

We then proceed as we did in Box 5.3, where we allowed mutation to occur. In this case, with migration, if either of the two alleles is a migrant, then they cannot be identical by descent (we ignore the remote possibility that the two alleles are both migrants and identical by descent). The probability that neither allele is a migrant is $(1 - m)^2$, so we, therefore, multiply the right side of the above equation by $(1 - m)^2$. We then solve this as we did in Box 5.3: we set $F_t = F_{t-1} = \widehat{F}$ (at equilibrium), and we assume m is sufficiently small that we can ignore terms with m^2 and set $1 - 2m = 1$. If you do all this, you end up with (following the steps in Box 5.3):

$$\widehat{F} = \frac{1}{4Nm + 1}$$

Q.E.D.

any particular population; over time, the allele frequency in each population will thus become identical to the average allele frequency over all populations. One can come up with fancier models, for example, allowing for different rates of migration between different populations, but the overall outcome will be the same: the allele frequency differences among populations will become smaller and smaller over time.

If we turn now to our two questions of interest, what does migration do in terms of genetic variation within populations and genetic differences among populations? Keep in mind our infinite alleles model of mutations: we have several populations, so the mutations occurring in each population are going to be different. Migration will spread these mutations around among the different populations, hence the genetic variation within each population will be increased (relative to what it would be without migration). And, it should be easy to see that genetic differences among populations are going to decrease with migration—over time, we expect the allele frequencies in the populations to converge to some average value.

Migration and Genetic Drift

As we saw previously, with small population sizes genetic drift will increase homozygosity, leading to an increase in genetic differences among populations. Migration, on the contrary, counteracts this effect of genetic drift by decreasing the genetic differences among populations. What happens when we have both genetic drift and migration: which one "wins"?

This situation turns out to be very similar in principle to the case discussed above of mutation versus genetic drift, and the relevant equation should look very familiar (with the derivation, for those of you interested in such details, in Box 5.5):

$$\widehat{F} = \frac{1}{4Nm + 1}$$

where $\widehat{F}$ is the usual inbreeding coefficient, that is, the probability that two alleles chosen at random are identical by descent. So, if $\widehat{F}$ is low, then migration has won (because migration is keeping homozygosity due to identity by descent low); if $\widehat{F}$ is high, genetic drift has won (because genetic drift is keeping homozygosity due to identity by descent high). A graph of the relationship between Nm and $\widehat{F}$ is shown in Figure 5.12. The take-home lesson from this graph is that it takes only a few migrants per generation to counteract the effects of genetic drift—and, remarkably, this number does not depend on the population size! Nm is the absolute number of migrants per generation, so one migrant per generation has the same effect when the population size is 1 million as when it is 100. This may seem very counterintuitive—surely migration should have a much bigger effect on the gene pool of a small population—but the way to think about

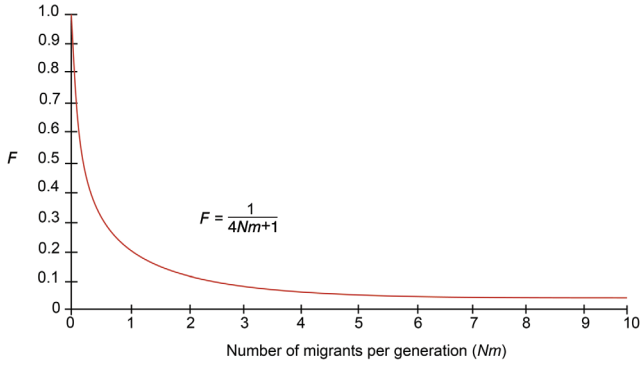


FIGURE 5.12

Relationship between the inbreeding coefficient (F) and the number of migrants per generation, Nm (where N is the effective population size and m is the migration rate). In this context, F can also be thought of as a measure of the differentiation among subpopulations (i.e., the bigger F is, the more different the subpopulations will be).

this is as follows: in a small population, a migrant does indeed have a bigger effect on the gene pool (in terms of contributing new alleles to the gene pool) but so does genetic drift (in terms of increasing homozygosity). In a large population, a migrant has a much smaller effect on the gene pool, but genetic drift is correspondingly much weaker in large populations. What is remarkable is that these two forces balance each other exactly, so population size does not matter.

Wahlund's Effect

There is one more topic to consider with respect to migration and that is what happens when what we think is one random-mating population is in fact two (or more) subpopulations that are not mating at random. Consider a simple example: suppose we have two subpopulations that do not mate with each other at all, of equal size, and one subpopulation is homozygous for the A allele and the other is homozygous for the B allele. Then allele frequencies in the total population are $p = q = 0.5$, so our expected heterozygosity (assuming Hardy–Weinberg equilibrium) is 0.5. Yet, the observed heterozygosity is 0. It turns out that whenever there is hidden population structure (i.e., subpopulations that we don't know about that are not mating at random with each other, even though each subpopulation is at Hardy–Weinberg equilibrium), the observed heterozygosity will always be less than the expected heterozygosity. This reduction in heterozygosity with population structure is known as **Wahlund's effect**, after the Swedish geneticist Sten Wahlund, who was the first to document this effect (Wahlund 1928). We will see examples of population substructure in more detail later (e.g., Figure 11.18), as it is often the case with human populations that there is substructure, that is, subgroups which are not mating at random.

An important corollary of the Wahlund effect is what happens when formerly isolated subpopulations start exchanging migrants (and hence, alleles). Here, we expect homozygosity to decrease and heterozygosity to increase (if this is not obvious, go back to the simple aforementioned example of two subpopulations fixed for different alleles—what happens if they now start mating randomly?). This in turn means that the incidence of recessive diseases (i.e., those caused by homozygosity for a recessive allele) should also decrease. Thus, as the trend these days is toward the reduction of population isolation due to increased local and global mobility, we can expect the incidence of recessive diseases to decrease—which is good news indeed.

SELECTION

The final evolutionary force to consider is the one that probably occurs to most people when they think of evolution, and that is natural selection, or just selection for short. Selection is a key aspect of Darwinian evolution, which simply stated, holds that given:

- variation among individuals in their ability to procure and utilize resources,
- this variation is at least partly transmitted from parents to offspring,
- limited resources (food, territory, access to mates, etc.) and thus competition between individuals for access to such limited resources,

then those characteristics that enhance an individual's chances of surviving and/or reproducing will increase in frequency from generation to generation. Note that all of these must hold for evolution via selection to occur: if there is no variation, or if the variation is not inherited, or if there is no competition, then there will be no selection. Furthermore, note that if all of these hold, then evolution via selection will necessarily occur.

So how does selection influence the genetic structure of populations? Although selection acts on phenotypes, which are in turn influenced by both the entire genome and the environment, we will focus on how selection influences the allele frequencies at a single gene. The model we will set up assumes that selection operates on **viability** (the chance of living to reproductive age) and can be diagrammed as follows:

Adults → Gametes → Random mating → Zygotes
→ Selection → Adults

According to this model, adults produce gametes, which then undergo random mating to produce zygotes, after which selection happens on the zygotes, and then we have the adults who will then produce the next round of gametes. For those of you keeping score, note that this definition of viability appears to differ from that in Chapter 4, where we defined viability as the probability of living until reproduction ceases. In fact, the definitions are the same, because in our aforementioned model we assume that once you reach reproductive age, you immediately reproduce and that is the end of the matter, so the beginning of reproductive age coincides with the end of reproductive age. We could also include selection on fertility, which of course is very important, but then the math gets much messier, because fertility is a property of a mating—your fertility depends not just on your own

genotype but also on the genotype of your mate. Your genotype could mean that you are extremely fertile, but if your mate is sterile, well, too bad. Viability, on the contrary, is a property of an individual, so the math is easier to deal with, and the overall principles are the same.

We start with our usual gene pool model of two alleles, A and B, with allele frequencies p and q, respectively. We now add a new parameter, W, which we define as the relative chance of surviving to reproduce, also known as the **fitness**. Do not confuse this concept of fitness with physical fitness, as all evolution cares about is the production of offspring (who in turn survive to produce offspring, etc.). If you spend all your time working out you may be quite physically fit, but if all that recreation leaves you too tired for procreation, whereas your couch potato friend who never exercises ends up with lots of children, then your friend is fitter than you!

We can set up our gene pool model that incorporates selection as follows: we have our usual two alleles, A and B, with frequencies p and q, respectively. Then with selection:

Genotype:	AA	AB	BB	Total
Frequency:	p^2	$2pq$	q^2	1
Fitness:	W_{AA}	W_{AB}	W_{BB}	
Relative contribution:	p^2W_{AA}	$2pqW_{AB}$	q^2W_{BB}	$\overline{W}$
Normalized frequency:	$\frac{p^2W_{AA}}{\overline{W}}$	$\frac{2pqW_{AB}}{\overline{W}}$	$\frac{q^2W_{BB}}{\overline{W}}$	1

The idea is that we start with our genotypes in Hardy–Weinberg equilibrium in the offspring. Each genotype has an associated fitness value, which is where selection enters the picture: multiply the fitness value by the genotype frequency to get the relative contribution of that genotype to the adults. We define the **average fitness** of the population, $\overline{W}$, to be the sum of the relative contributions. Divide each relative contribution by the average fitness to get the normalized frequency of each genotype in the adults, which can then be used to calculate p and q in the adults, which can then be used to calculate the genotype frequencies in the next generation of offspring (using Hardy–Weinberg).

Let's work through an example. Suppose we start with allele frequencies of $p = q = 0.5$, and fitness values of 4, 3, and 2 for the AA, AB, and BB genotypes, respectively. These fitness values can be thought of as the relative number of each genotype that survives to reproductive age, that is, for every 4 AA individuals who survive to reproduce, 3 AB and 2 BB individuals also survive to reproduce. Then with the aforementioned model:

Genotype:	AA	AB	BB	Total
Frequency:	0.25	0.5	0.25	1
Fitness:	4	3	2	
Relative contribution:	1	1.5	0.5	3
Normalized frequency:	0.33	0.5	0.17	1

After selection has happened, the frequency of AA homozygotes has increased and that of BB homozygotes has decreased, while the frequency of AB heterozygotes is unchanged. What are p and q in the adults? It turns out that $p = 0.58$ and $q = 0.42$ (go back to the beginning of Chapter 4 if you don't remember how to do this). So, selection has increased the frequency of the A allele from 0.5 to 0.58 and correspondingly decreased the frequency of the B allele to 0.42. And what happens in the next generation? I leave that as an exercise for you to do; use Hardy–Weinberg with $p = 0.58$ and $q = 0.42$ to get the genotype frequencies in the offspring in the next generation and assume that there is no change in the fitness values. You should then end up with $\overline{W} = 3.16$, genotype frequencies of 0.43, 0.46, and 0.11 for the AA, AB, and BB genotypes, respectively, and $p = 0.66$ and $q = 0.34$.

So in this example, selection is increasing the frequency of the A allele. Moreover, the average fitness is also increasing each generation—it went up from 3 to 3.16 after one generation. What do you suppose will eventually happen if selection continues with these fitness values? Hopefully, it is fairly obvious that eventually the A allele will become fixed and the B allele lost, and when that happens the average fitness of the population will be the maximum possible value of 4 (because $p = 1$). The take-home lesson: selection always operates to increase the average fitness of a population—at least, in the simple models we will be concerned with—and so allele frequencies will change over time accordingly.

This is an example of **directional selection**, in which the fitness of one homozygous genotype is higher than that of the other, and the heterozygote has either intermediate fitness (as in the aforementioned case) or fitness equal to one of the homozygotes. In principle, we can speak of either positive directional selection (or just **positive selection**, for short), in which selection results in an increase in the frequency of a favorable allele, or **negative selection**, in which selection results in a decrease in the frequency of an unfavorable allele. In reality, these are just two sides of the same coin; by definition, if selection is increasing the frequency of an allele at a locus because it has

the highest fitness, then selection is simultaneously decreasing the frequency of all other alleles at that locus because they are less favorable than the selected allele.

With directional selection, we expect to ultimately get fixation of the allele with the highest associated fitness. How long does this take? While we could use the framework above to figure this out, the algebra is rather messy and it is not so easy to see what is going on. So we will instead introduce a different framework that is easier to interpret, in which we set the fitness values to the following:

$$W_{AA} = 1$$
$$W_{AB} = 1 - hs$$
$$W_{BB} = 1 - s$$

In this framework, s is the **selection coefficient** and h is the **degree of dominance**. When $s = 0$, there is no selection; when $s > 0$, then there is selection against the B allele (and for the A allele); and when $s < 0$, there is selection against the A allele (and for the B allele). And when $h = 0$, the A allele is completely dominant with respect to the B allele; when $h = 1$, the B allele is completely dominant with respect to the A allele; and when $0 < h < 1$, there is intermediate dominance (we will consider later what happens if $h < 0$ or $h > 1$). It is easy enough to translate any set of fitness values into this framework and figure out what h and s are. For example, with our previous values of $W_{AA} = 4$, $W_{AB} = 3$, and $W_{BB} = 2$, we first divide all values by 4 to get $W_{AA} = 1$; then $W_{AB} = 0.75$, and $W_{BB} = 0.5$. From these numbers, we can work out that $s = 0.5$ (from the equation for W_{BB}) and $h = 0.5$ (from the value for s and the equation for W_{AB}).

Now we are ready to address the question as to how long it takes for fixation of a favorable allele (or loss of an unfavorable allele) to occur. An example is shown in Figure 5.13; the length of time to get fixation/loss depends on both the strength of selection and the degree of dominance (i.e., both s and h are important). In general, the stronger the selection (in terms of the difference in fitness values between the two homozygous genotypes), the faster fixation of the favored allele occurs, which hopefully makes sense. But also note the big difference in the dynamics of the process between a dominant and a recessive allele. When a new mutation occurs that is both dominant and favored, it initially increases in frequency quickly, but then the rate of change slows drastically, whereas a new mutation that is both recessive and favored initially increases very slowly, until at a certain point the rate of change increases dramatically. Why this difference? Well, keep in mind that when a new mutation occurs, it will initially be present only in heterozygotes. For a new dominant mutation, this doesn't matter, because by definition the heterozygote will exhibit the associated phenotype, hence selection will immediately start acting upon the new mutation and increasing its frequency. But as the favorable allele increases in frequency, there will be fewer and fewer homozygotes for the unfavorable allele—and heterozygotes (having both the favorable and the unfavorable alleles) have, by definition, the same fitness as homozygotes for the favorable allele. Therefore, selection becomes less efficient at eliminating the unfavorable allele when it becomes rare. You can check for yourself: assuming Hardy–Weinberg, an allele with a frequency of 0.1 has an 18:1 ratio of heterozygotes to homozygotes (i.e., there are 18 heterozygotes for every homozygote), while an allele with a frequency of 0.01 has a ratio of 198:1 of heterozygotes to homozygotes. So, the rarer the allele, the more likely you are to find heterozygotes rather than homozygotes for the allele. And when a new, favorable mutation arises that is recessive, initially it is present mostly in heterozygotes, so it takes a long time for the allele frequency to rise to the point that there are enough homozygotes for their fitness advantage to become apparent and thus for selection to become efficient at raising the frequency of the favorable mutation.

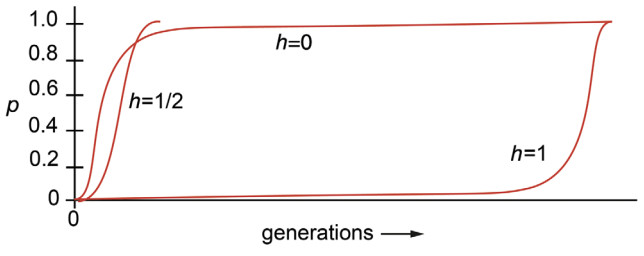


FIGURE 5.13

Change in allele frequency (p) over time with positive selection when the allele is dominant ($h = 0$), recessive ($h = 1$), or partially dominant ($h = 0.5$).

Mutation–Selection Balance

How important is directional selection in humans? The importance of positive selection is currently a matter of some controversy, as discussed in Chapter 20. However, negative selection is extremely important, as there are numerous examples known of diseases caused by genetic mutations that decrease viability and/or fertility. In fact, it is estimated that about 5% of all newborns are afflicted with a genetic disease; genetic diseases, therefore, have an important impact on human health that—especially as more and more vaccines or cures are found for the infectious diseases that plague us—continues to grow. So why hasn't selection eliminated all of these unfavorable, disease-causing alleles? The answer is because while selection

BOX 5.6 ■ Mutation–Selection Balance

Start with our framework for selection, with

$$W_{AA} = 1$$
$$W_{AB} = 1 - hs$$
$$W_{BB} = 1 - s$$

Then we have:

Genotype	AA	AB	BB	Total
Frequency	p^2	$2pq$	q^2	1
Fitness:	1	$1 - hs$	$1 - s$	
Relative contribution:	p^2	$2pq(1 - hs)$	$q^2(1 - s)$	$\overline{W}$
Normalized frequency	$\frac{p^2}{\overline{W}}$	$\frac{2pq(1 - hs)}{\overline{W}}$	$\frac{q^2(1 - s)}{\overline{W}}$	1

Let's consider first the case where the B allele is completely recessive with respect to the A allele; then $h = 0$ and we have:

$$\overline{W} = p^2 + 2pq + q^2(1 - s)$$
$$= 1 - sq^2 \text{ (because } p^2 + 2pq + q^2 = 1)$$

In the next generation, p' is gotten by taking the frequency of AA homozygotes plus half the frequency of AB heterozygotes. However, we now include the possibility of mutation from an A allele to a B allele at a rate μ (we ignore the possibility of mutation from a B allele to an A allele, which is justifiable if, for example, the B allele is a disease allele and hence expected to be rare). Therefore, the A alleles in the next generation will be reduced by a factor μ. So,

$$p' = \frac{p^2 + pq}{\overline{W}}(1 - \mu) \quad \text{note that } p^2 + pq = p(p + q) = p$$
$$= \frac{p(1 - \mu)}{1 - sq^2}$$

And the change in the frequency of the A allele from one generation to the next is:

$$p' - p = \frac{p(1 - \mu)}{1 - sq^2} - p$$

By definition, at equilibrium the gain in B alleles by mutation is balanced by the loss of B alleles due to selection, hence $p' - p = 0$:

$0 = \frac{p(1 - \mu)}{1 - sq^2} - p$ add p to both sides, divide by p, and multiply by $(1 - sq^2)$

$1 - sq^2 = 1 - \mu$ subtract 1 from both sides and divide by $(-s)$

$q^2 = \mu/s$

Q.E.D.

Next, let's consider the case of partial dominance, so $h > 0$. Then we have:

$$\overline{W} = p^2 + 2pq(1 - hs) + q^2(1 - s)$$
$$= 1 - 2pqhs - sq^2$$

and when we add mutation from A to B by the same aforementioned reasoning:

$p' = \frac{p^2 + pq(1 - hs)}{\overline{W}}(1 - \mu)$ note that $p^2 + pq = p$, substitute this and multiply everything out

$$= \frac{p - pqhs - \mu p + \mu pqhs}{1 - 2pqhs - sq^2}$$

At equilibrium, $p' = p$, so subtract p from both sides and set $p' - p = 0$:

$0 = \frac{p - pqhs - \mu p + \mu pqhs}{1 - 2pqhs - sq^2} - p$ add p to both sides, divide by p, and multiply by the denominator

$1 - 2pqhs - sq^2 = 1 - qhs - \mu + \mu qhs$ subtract 1 from both sides and substitute $(1 - q)$ for p

$-2qhs + 2q^2hs - sq^2 = -qhs - \mu + \mu qhs$

This might look rather hopeless, but keep in mind that μ is expected to be pretty small (no bigger than, say, 10^{-5}) and q is also expected to be small (much less than 0.01), so we can safely assume that terms with μq or q^2 in them can be set to 0, which give us:

$$-2qhs = -qhs - \mu$$

Solve this for q, and you should end up with:

$$q = \mu/hs$$

Q.E.D.

is eliminating unfavorable alleles, new copies are being created by mutation, leading to a balance between mutation and selection.

We'll consider two cases of mutation–selection balance. Box 5.6 goes through the details for those who are interested; here we'll just present the equations. The first case has to do with selection against a completely recessive allele (so $h = 0$); at equilibrium

$$q^2 = \mu/s$$

where μ is the mutation rate from the favorable (A) allele to the deleterious (B) allele. The general rule is that the frequency of a recessive deleterious trait is expected to be the ratio of the mutation rate to the selective disadvantage of the trait. Note that in the special case of a recessive lethal trait (meaning that homozygotes either die before reaching reproductive age or do not reproduce), $s = 1$ and $q^2 = \mu$; the frequency of a recessive lethal trait is just the mutation rate.

This suggests that we can use this fact to estimate the mutation rate for a recessive lethal disease. For example, phenylketonuria (PKU) is a recessive disease in which affected individuals have mutations that destroy their ability to metabolize the amino acid phenylalanine. Left untreated, phenylalanine builds up to high levels in the bloodstream, causing brain damage and mental retardation. Phenylketonuria is not, strictly speaking, a recessive lethal disease, but people with untreated PKU tend to be so severely affected that their reproduction is greatly depressed, so for our purposes, we can think of it as recessive lethal. The incidence of PKU is about 1 in 25,000 births in the United States, so our estimate of q^2 is $1/25{,}000 = 4 \times 10^{-5}$, which is also then our estimate of the mutation rate. This is higher than the usual mutation rates of around 10^{-8}, but keep in mind that these are mutation rates per DNA nucleotide. Here, we are concerned with all mutations that result in PKU, so any mutation that reduces/destroys the function of the enzyme that metabolizes phenylalanine will contribute to the mutation rate estimate. Incidentally, PKU can be treated with a special diet that reduces phenylalanine intake (but does not eliminate it completely, as some phenylalanine is a necessary component of our proteins). It used to be thought that this special diet was necessary only up until about age 6 years, when brain growth and development is largely complete, but subsequent studies documented further adverse effects when the diet was discontinued, so now individuals with PKU generally stay on the special diet for their entire lifetime. It is particularly important for women with PKU to be on the diet during pregnancy, as otherwise the high levels of phenylalanine in their bloodstream can damage the brain of the fetus (even though the fetus does not have PKU)—sadly, this was realized only after many unfortunate women with PKU, who had discontinued the special diet in childhood since that was prevailing medical wisdom at the time, subsequently had babies with brain damage.

In the case of partial dominance ($h > 0$), the allele frequency when there is an equilibrium between selection and mutation can be approximated as (Box 5.6):

$$q = \mu/hs$$

Suppose the heterozygote is exactly intermediate in fitness between the two homozygous genotypes, then $h = 0.5$ and $q = 2\mu/s$. How does this compare to selection against a completely recessive allele? Let's go back to our PKU example; with a mutation rate of 4×10^{-5}, and $s = 1$, then at equilibrium the frequency of the disease allele would be about 0.0063. But if PKU were instead semidominant with $h = 0.5$ and this same mutation rate, then at equilibrium the frequency of the disease allele would be much lower, only about 0.00008. Why is this? Think back to the results in Figure 5.13—the reduced fitness of heterozygotes in the semidominant case means that selection would be more efficient at eliminating the deleterious PKU allele, thereby driving down the equilibrium allele frequency. But in reality, since heterozygotes for PKU have the same fitness as homozygotes for the non-PKU allele, selection can eliminate only PKU alleles in PKU homozygotes. The end result is a higher equilibrium frequency for deleterious recessive alleles, compared to deleterious partially dominant alleles.

Balancing Selection

So far we have been considering directional selection, in which one of the homozygous genotypes has the highest fitness and the other homozygous genotype has the lowest fitness. What happens if the heterozygous genotype has the highest fitness? According to our model ($W_{AA} = 1$, $W_{AB} = 1 - hs$, $W_{BB} = 1 - s$), this can happen if $s > 0$ and $h < 0$ (in which case $W_{AA} > W_{BB}$), or if $s < 0$ and $h > 0$ (in which case $W_{BB} > W_{AA}$). It turns out that if the heterozygote has the highest fitness, then there is an equilibrium—when allele frequencies are no longer changing—for some intermediate value of p. The relevant equation (with details in Box 5.7 for those who are interested) is:

$$\hat{p} = \frac{(1-h)}{(1-2h)}$$

Note that the equilibrium frequency, $\hat{p}$, depends only on the degree of dominance (h) and not at all on the strength of the selection (s). A schematic depiction of p versus $\overline{w}$ appears in Figure 5.14 and hopefully makes clear why $\hat{p}$ is a stable equilibrium: the average fitness for the population is at a maximum when $p = \hat{p}$. Recall that selection always increases the average fitness, so if $p < \hat{p}$, p will increase until it reaches $\hat{p}$. And, if $p > \hat{p}$, p will decrease until it reaches $\hat{p}$.

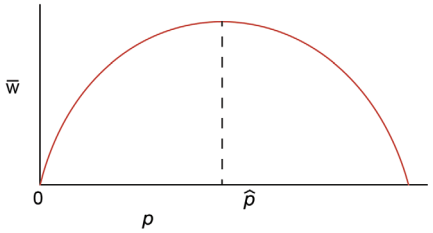


FIGURE 5.14

Relationship between average fitness ($\overline{w}$) and allele frequency (p) with balancing selection. At the equilibrium value of p ($\hat{p}$), average fitness is maximized and hence this is a stable equilibrium.

BOX 5.7 ■ Equilibrium Values with Selection

Let's go back to our model for selection (in the absence of mutation) and consider the change in allele frequencies over time. As was shown in Box 5.6, we can write the change in allele frequencies from one generation to the next as:

$$p' - p = \frac{p^2 + pq(1 - hs)}{\overline{W}} - p$$

Recall from Box 5.6 that $\overline{W} = 1 - 2pqhs - sq^2$, so get everything on the right side over the denominator and multiply to get rid of the parentheses:

$$p' - p = \frac{p^2 + pq - pqhs - p + 2p^2qhs + spq^2}{\overline{W}}$$

recall that $p^2 + pq = p$

$$= \frac{-pqhs + 2p^2qhs + spq^2}{\overline{W}}$$ factor out spq

$$= \frac{spq(-h + 2ph + q)}{\overline{W}}$$ note that $2ph = ph + (1 - q)h = ph + h - qh$

$$= \frac{spq(-h + ph + h - qh + q)}{\overline{W}}$$ simplify to

$$p' - p = \frac{spq[ph + q(1 - h)]}{\overline{W}}$$

We can now ask, when will the allele frequencies stop changing over time? That will happen when $p' - p = 0$, which in turn happens in three boring cases and one interesting case:

1. when $s = 0$ (no selection, so no change in allele frequencies because of selection)
2. when $p = 0$ (fixation of the *B* allele, so no more variation)
3. when $q = 0$ (fixation of the *A* allele, so no more variation)
4. when $[ph + q(1 - h)] = 0$ (which is interesting!)

Why is this last case interesting? Because it implies that there is some intermediate allele frequency that will not change over time because of selection. What is this intermediate frequency? We have

$$ph + q(1 - h) = 0$$

Set $q = 1 - p$ and work through the algebra (which, if you've been trying to follow the previous boxes, should be pretty easy for you!), you should end up with

$$\hat{p} = \frac{1 - h}{1 - 2h}$$

(Note that this is the same as $\frac{h-1}{2h-1}$)

Q.E.D.

This type of selection is known as **balancing selection** and is sometimes also referred to as **heterozygote superiority** because the heterozygous genotype has the highest fitness. The classic example of balancing selection in humans is the genetic disease sickle-cell anemia. The disease is caused by a recessive allele; homozygotes for the disease allele, denoted *S*, have red blood cells that become distorted and adopt a sickle (curved) shape. This in turn decreases their elasticity and accelerates their destruction, leading to anemia and other complications. By the 1940s, it was well known that the disease occurred primarily in people of African ancestry, and, moreover, that it had very high mortality—most victims died before reaching puberty. Yet, it was puzzling why a disease with such high mortality should also occur at such high frequency in some African populations. The puzzle was solved in a series of papers in the 1950s by Anthony Allison (Allison 1954a, 1954b), who noticed that there was a strong overlap in the distribution of sickle-cell anemia and malaria and postulated that heterozygotes for the sickle-cell anemia allele (genotype *AS*) were resistant to malaria. He later showed that children in areas where malaria was endemic had lower malaria parasite counts if they were heterozygous *AS* than if they were homozygous *AA*, and that there was a reduced frequency of *AS* heterozygotes among fatalities due to malaria (Allison 1956, 1957).

We can use Allison's original data to estimate the equilibrium frequency for the sickle-cell allele. He estimated the relative viabilities of the genotypes as $AA = 0.85$, $AS = 1.0$, and $SS = 0$. If you turn these into fitness values and plug them into our formulas for h and s (remember to first divide so as to have the fitness of the *AA* genotype = 1), you should get that $s = 1$ (so this is a recessive lethal disease—at least it was in Africa in the middle of the twentieth century) and $h = -0.18$. The equilibrium frequency of the *S* allele is then, from the previous equation, $\widehat{q} = 1 - \hat{p} = 0.13$. In actuality, the observed frequency of the *S* allele in the populations studied by Allison is about 0.09, which is a pretty good fit considering all the assumptions that go into this calculation, suggesting that the sickle-cell anemia allele is near or at equilibrium.

It turns out that the sickle-cell anemia allele is due to a mutation in the gene for β-globin, which is one of two types of polypeptide chains that make up the adult form of hemoglobin. Hemoglobin is the major protein in red blood cells, which are responsible for transporting oxygen throughout the body. This was the

first demonstration of a link between a specific disease and a mutant protein and was carried out by Linus Pauling (who later went on to win two Nobel Prizes, one in chemistry for his work on chemical bonds, the other the Nobel Peace Prize for his efforts to end nuclear weapons testing) and colleagues (Pauling et al. 1949). Pauling referred to this as the first "molecular disease" and speculated that sickle-cell anemia might represent an intermediate stage in the evolution of resistance to malaria. He also became interested in **eugenics**, the use of genetics to "improve" the human species, and advocated that people should be tested for the sickle-cell allele. Pauling thought that carriers (heterozygotes) should be prevented from having children with each other as 25% of their offspring would have sickle-cell anemia, saying "This percentage [25%] is much too high to let private enterprise in love combined with ignorance take care of the matter" (Zuckerkandl and Pauling 1962). He even went so far as to suggest that

> … the time might come in the future when information about heterozygosity in such serious genes as the sickle cell anemia gene would be tattooed on the forehead of the carriers, so that young men and women would at once be warned not to fall in love with each other. (Pauling 1966)

So, just because somebody is smart enough to win a Nobel Prize or two doesn't mean that he or she isn't capable of rather questionable ideas on other subjects!

Disruptive Selection

For the sake of completeness—and also because it is pretty weird!—let's also consider the case where the heterozygous genotype has the lowest fitness. Known as **disruptive selection** or **heterozygote inferiority**, this occurs when either $h > 1$ and $s > 0$, or $h < 0$ and $s < 0$. It turns out that under these conditions there is also an intermediate equilibrium frequency for p:

$$\hat{p} = \frac{(1-h)}{(1-2h)}$$

This is the same equilibrium frequency for p as in the case of balancing selection, but with one important exception: with balancing selection $\hat{p}$ is a stable equilibrium, whereas with disruptive selection $\hat{p}$ is an unstable equilibrium, as shown in Figure 5.15, which plots p versus $\overline{W}$. Note that $\overline{W}$ is at a minimum when $p = \hat{p}$, but this is an equilibrium point (in the sense that p will not change because of selection when $p = \hat{p}$). However, as soon as anything happens to move p from this equilibrium point (such as random changes due to genetic drift, migration, etc.), selection will operate to move p

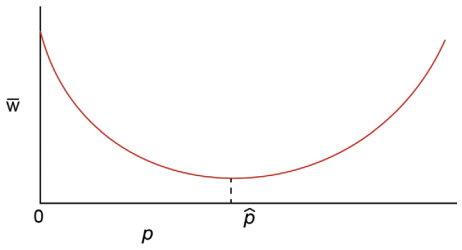


FIGURE 5.15

Relationship between average fitness ($\overline{w}$) and allele frequency (p) with disruptive selection. At the equilibrium value of p ($\hat{p}$), average fitness is minimized and hence this is an unstable equilibrium.

away from $\hat{p}$ and not toward it as in the case of balancing selection. Thus, with disruptive selection we expect fixation of one allele or the other, and not maintenance of the intermediate equilibrium value, $\hat{p}$. Pretty strange, no?

Examples of disruptive selection are quite rare; probably the best-known example involves a particular type of chromosomal rearrangement known as a **translocation**. This involves the exchange of parts of two chromosomes between one another (see Figure 5.16). Assuming that the actual break in the DNA sequence of the chromosomes does not disrupt any important function, then a heterozygote for such a translocation has the same DNA as either homozygote, just rearranged differently, and so is perfectly normal. But during meiosis (the production of gametes), the normal chromosomal pairing is disrupted in a heterozygote for a chromosome translocation, and the result is that about 50% of the gametes will be unbalanced, having duplications and deficiencies in their DNA content (Figure 5.16). Since many genes

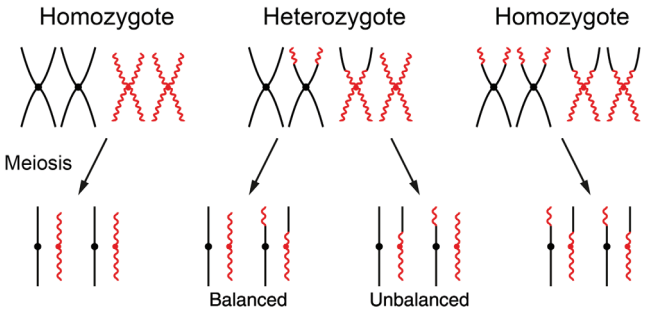


FIGURE 5.16

Products of meiosis from a reciprocal chromosomal translocation. Homozygotes produce gametes with the full complement of genes, while heterozygotes produce some gametes that carry duplications and deficiencies of chromosomal segments; these are likely to be deleterious. Hence, chromosomal translocations are an example of disruptive selection.

are usually influenced by these duplications and deficiencies, the resulting offspring will in most cases die before birth (chromosomal abnormalities are frequently implicated in spontaneous abortions) and even if born alive will suffer profound and severe birth defects. Thus, translocation heterozygotes have a reduction in fertility, whereas homozygotes for either the translocated or the untranslocated chromosomes have no such problems with chromosome pairing during meiosis, and hence no fertility reduction. And while such chromosomal translocations are rare events, they nonetheless have occurred during human evolution. For example, humans differ from other apes by having 23 rather than 24 pairs of chromosomes, and it is quite clear that this reduction in chromosome number occurred by the fusion of two chromosomes sometime after our lineage diverged from that of apes. Even though there would (presumably) have been disruptive selection against heterozygotes for this chromosomal rearrangement, in this case the altered chromosome arrangement still managed to increase in frequency and become fixed in our lineage.

Selection: Summary

Having discussed various aspects of selection, we are now ready to ask about the effect of selection on genetic variation within populations and genetic differences among populations. The effect of selection on genetic variation is easy: directional selection results in the fixation of favorable alleles and loss of unfavorable alleles, and hence loss of genetic variation; balancing selection results in an intermediate allele frequency and hence maintenance of genetic variation (we will ignore disruptive selection but it shouldn't be hard for you to figure out what disruptive selection does to genetic variation). But what effect does selection have on genetic differences between populations—does it tend to increase or decrease genetic differences? The answer is, it depends on whether the reason for selection is the same or different between populations. If two populations are experiencing more or less the same environment—specifically, the same challenges to survival—then we expect selection (either directional or balancing) to behave similarly in the two populations and hence to decrease the genetic differences between the populations. But if two populations differ with respect to some important feature of the environment—in particular, they have to cope with different challenges in order to survive and reproduce—then selection may operate differently in the two populations and hence increase the genetic differences between them.

For example, take the case of sickle-cell anemia. In a malarial environment, there is balancing selection for the sickle-cell allele, and in African populations where malaria is endemic, there are similar frequencies of the sickle-cell allele. Same environment, same selective pressure, and the result is decreased genetic differences between populations. But in the case of African–Americans, there is no malaria in the United States, so no balancing selection. Instead, African–Americans are experiencing directional selection against the sickle-cell allele, and so selection is increasing the genetic differences between African–American and native African populations.

It is important to distinguish between the effects of selection at the phenotypic versus molecular genetic levels, as different events at the genetic level can have similar phenotypic consequences. For example, malaria is endemic not just in parts of Africa but also in many other parts of the world, including Southeast Asia and Oceania. Human populations in all of these regions have evolved some form of genetic resistance to malaria, but different mutations have been selected for in different populations, with different consequences for human health. This is an example of **convergent evolution**, whereby different mutations at the molecular level give rise to the same phenotype (more or less) and hence are subject to the same selection. Convergent evolution is quite common; another example that will be discussed in detail in Chapter 18 is **lactose tolerance**, the ability to digest milk into adulthood. The fact that convergent evolution happens so frequently is an indication that there tend to be many different ways (at the molecular level) of responding to a particular "problem" imposed by the environment (e.g., malaria or digesting milk).

There is one other aspect of selection to be discussed and that is the role of selection versus genetic drift in determining the fate of new favorable mutations. When a new mutation arises, it is present by definition in just one copy in the population. Suppose this mutation is recessive and that homozygotes for this mutation have a 10% increase in fitness (which is a huge increase!). What do you suppose will happen to this mutation? In an infinite population, this favorable mutation will inexorably rise in frequency and eventually sweep through the population to fixation, but in real life populations aren't infinite in size. It is easiest to see what happens by using one of the simulation programs such as Allele A1. I've just done this 10 times each for populations of size 50, 500, and 5000 (changing the initial allele frequency to 0.01, 0.001, and 0.0001, respectively, to correspond to a new mutation); in each case, the favorable new mutation was *lost* in all 10 simulations. A moment's thought should make clear why this is the case: until the allele frequency becomes high enough to produce homozygotes for the favorable allele, selection will not have a chance to operate. Thus, genetic drift will govern the fate of the new favorable mutation, and as with any low

TABLE 5.1 ■ Summary of the influence of each evolutionary force on genetic variation within populations and genetic differences among populations

Evolutionary force	Genetic variation within populations	Genetic differences among populations
Inbreeding	Decreases	Increase
Small population size (drift)	Decreases	Increase
Mutation	Increases	Increase
Migration	Increases	Decrease
Selection (directional)	Decreases	Same environment: decrease Different environment: increase
Selection (balancing)	Increases	Same as for directional selection

frequency allele, most of the time the new favorable mutation will be lost due to genetic drift. And changing the effect of the favorable mutation to be completely dominant (and still with a huge fitness advantage of 10%) doesn't help much; now fixation of the favorable mutation happened in just 1–4 of each set of 10 simulations. The sobering conclusion: many potentially advantageous mutations that occurred during our evolution never got a chance to show what they could do, as they were quickly eliminated by genetic drift. And those advantageous mutations that eventually did rise in frequency and become fixed in our species were the lucky ones that first by genetic drift—and not by their selective advantage—rose to a high enough frequency that selection could then act on them. It does make you wonder what humans would be like now if more (or different) selectively advantageous alleles had made it through the filter of genetic drift.

EVOLUTIONARY FORCES: SUMMARY

This chapter has covered a lot of material that involves a lot of algebra. However, I have tried to emphasize the logic behind the various evolutionary forces, with the mathematical details presented in boxes for those who are interested in the gory details. Table 5.1 summarizes the important points from this chapter, namely, the expected impact of each evolutionary force on genetic variation within populations and genetic differences between populations; go back to the pertinent section if something in the table is not familiar or understandable. The point of going over all of this material is to have an understanding of the potential explanations for the various patterns of genetic variation within populations, and genetic differences among populations, that we will encounter later.

CHAPTER 6

MOLECULAR EVOLUTION

The previous chapter covered how the gene pool changes in response to various evolutionary forces, including inbreeding, small population size, mutation, migration, and selection. All of the results discussed in this chapter were derived long before anybody knew what a gene actually was. That is why many of the examples that were presented refer to genetic diseases such as sickle-cell anemia or PKU, because for a long time these were the only traits that geneticists had to work with. But we now know that genes consist of DNA, and it turns out that—in addition to the evolutionary forces discussed previously—there are other important aspects concerning how DNA and genes evolve that are related to their molecular nature. So, in this chapter, we will discuss these features of **molecular evolution**.

Following Darwin, for many years, the dominant paradigm in biology was evolution via Darwinian natural selection, which, to remind you, works as follows:

1. if there is variation among individuals in their ability to obtain resources, and these differing abilities influence their chances of surviving and reproducing (i.e., there is variation in fitness),
2. this variation is at least partly heritable (i.e., your fitness is at least partly related to the alleles you received from your parents),
3. and there is competition among individuals for such resources (i.e., there is not enough food, breeding space, potential mates, etc., for everyone),

then it is inevitable that those genetic variants that increase an individual's chances of surviving and reproducing will increase in frequency over time. Note that all three factors are necessary for evolution via natural selection to occur. If everyone has the same fitness, or if there is no genetic variation that influences fitness, or if resources are sufficiently abundant that everyone gets everything they need to survive and reproduce regardless of their individual fitness, then evolution via natural selection cannot occur. But also note that if all three factors are present, then by definition evolution via natural selection will occur.

If Darwinian evolution also holds at the molecular level, then genetic differences between species should largely reflect this process of adaptive evolution. That is, genetic differences observed between species should be meaningful in the sense that they should have had an important impact on fitness, thereby causing them to be selectively favored. However, as methods were developed for investigating variation at the molecular level—first with proteins, because they were more abundant and easier to work with than DNA, but later with DNA variation as well—it became clear that many aspects of molecular variation do not conform with expectations based on evolution via natural selection. Instead, it appears that most of the genetic differences between species have little if any effect on fitness. The Japanese geneticist Motoo Kimura was one of the first to notice this, and he put together an alternative explanation for molecular variation known as **neutral theory** (Kimura 1968, 1983). There are three main features to neutral theory, discussed in turn in the next three sections.

Before going into the evidence concerning how molecules (proteins and DNA) evolve, there is an important point of terminology to discuss, and that is the difference between **mutations** and **substitutions**. As we have seen, mutations are any changes that occur in a DNA sequence. Substitutions are mutations that have risen in frequency to become fixed within a species and hence constitute genetic differences between species. Mutations occur at random with respect to DNA sequences (more or less—some types of mutations are more common than others purely for biochemical reasons) and their associated phenotypic effects; the frequency with which a mutation occurs is not related to how benign or deleterious it is. Substitutions, on the contrary, are mutations that have survived the filter of natural selection to rise in

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

frequency and become fixed, with the harmful mutations weeded out. So keep in mind that when I refer to substitutions, these are the fixed genetic differences that are actually observed between species, not the complete spectrum of mutations that can potentially occur.

FUNCTIONALLY LESS IMPORTANT MOLECULES (OR PARTS OF MOLECULES) EVOLVE FASTER THAN MORE IMPORTANT ONES

If genetic differences between species are largely adaptive, then they should influence the function of the gene. Therefore, we should find genetic differences occurring more often in genes that encode functionally important proteins. We should also find genetic differences occurring more often in those parts of a protein that are important for function and less often in those parts of a protein that are not so important for function.

To test this prediction, we need to have a way to measure rates of molecular evolution. This was first done with proteins, for which rates of change are measured in units of amino acid substitutions per site per billion years. So, for a protein that evolves at a rate of 1, there is on average one amino acid substitution at each position in the protein's polypeptide chain every billion years—which just goes to show how slow evolution really is! Incidentally, the term "pauling" (for Linus Pauling) has been proposed to refer to molecular evolutionary rates to correspond to using "darwins" to measure rates of phenotypic evolution—one darwin is one unit change in a phenotype per billion years. So a protein with an evolutionary rate of 1 would evolve at a rate of 1 pauling, but this terminology never caught on. Evolutionary rates are estimated by determining the amino acid sequence of the protein's polypeptide chain from two (or more) species for which the fossil record (or other evidence) provides a reasonably good estimate of the species divergence time; we'll see in Chapter 12 how the rate estimation is actually done. Incidentally, determining the amino acid sequence of a protein used to be extremely laborious and time-consuming, involving purification of large quantities of the protein, followed by a complex series of biochemical manipulations to determine the amino acid sequence. In fact, when I was a graduate student in the late 1970s to early 1980s, determining the amino acid sequence of a single protein from a single species was so demanding that this would suffice for a PhD dissertation, and indeed some of my fellow graduate students were doing just that. Nowadays, DNA sequencing is much easier, faster, and cheaper than amino acid sequencing, so if you want to know the amino acid sequence of a particular protein, the way to do it is to simply sequence the gene for that protein and infer the amino acid sequence from the DNA sequence—and you won't get a PhD any more for doing this for just a single protein from a single species!

Anyway, as more and more amino acid sequences were determined for more and more proteins, it became clear that rates of molecular evolution are not correlated with the functional importance of molecules or parts of molecules. Instead, just the opposite seemed to hold—the less important a molecule (or part of a molecule), the faster the rate of evolution. For example, some of the fastest evolving proteins are fibrinopeptides, which have an evolutionary rate of about 8.3. Fibrinopeptides are formed during the blood clotting process, during which a soluble protein in the blood called fibrinogen is converted to insoluble fibrin, which provides the "scaffold" of the blood clot. The fibrinopeptides are part of fibrinogen but are released during the conversion process and do not play any further role in blood clot formation; they are ultimately degraded in the bloodstream. Thus, fibrinopeptides have no known function (other than to be released during blood clot formation), yet they evolve very rapidly, presumably precisely because virtually any amino acid substitution will still allow them to be released. Conversely, some of the slowest evolving proteins are histones, which have an evolutionary rate of about 0.008 (i.e., about 1000 times slower than fibrinopeptides). Histones bind to DNA and are an important component of the structure of chromosomes. Apparently, mutations that alter the amino acid sequence of histones are likely to make chromosomes less stable, which is very bad news indeed.

Another example is hemoglobin—recall that hemoglobin is a protein found in blood that binds to oxygen and transports it via the bloodstream throughout the body. Part of each hemoglobin polypeptide chain forms the surface of the hemoglobin molecule, and part forms the heme pocket, which actually binds to oxygen and hence is the functionally important part of hemoglobin. Yet, the rate of evolution of amino acids that make up the heme pocket is about 10 times less than the rate of evolution of amino acids on the surface of the protein. So, there is a slower rate of molecular evolution of the functionally important heme pocket than of the less important surface of hemoglobin.

What holds for proteins also holds for genes. Figure 6.1 illustrates the various parts of a gene and their approximate rates of molecular evolution—remember, a gene is composed of exons (the actual DNA sequence that encodes the protein product of the gene) and introns (noncoding DNA that interrupts the coding sequence), with noncoding DNA in between genes. In general, the exons evolve at a slower rate than the

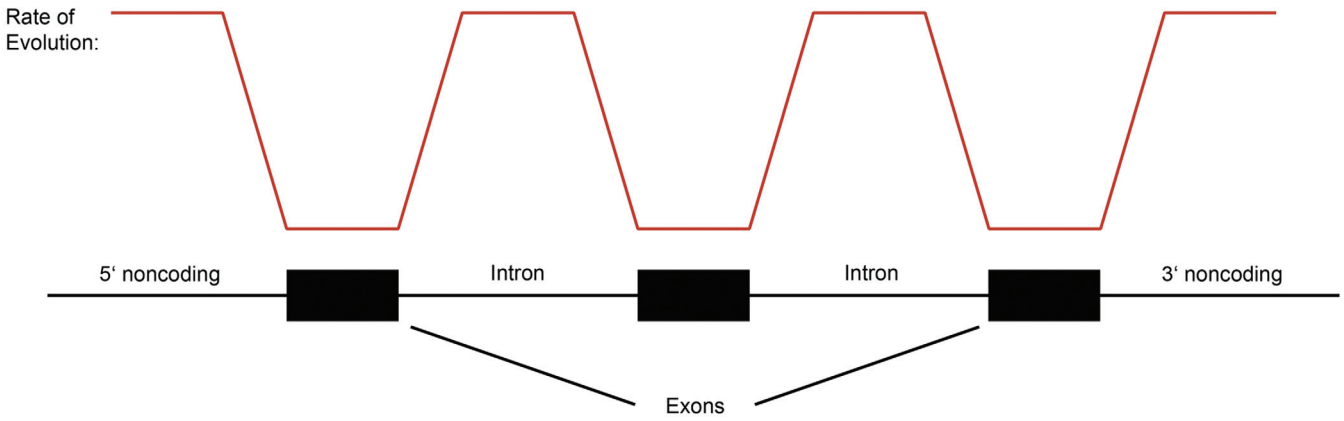


FIGURE 6.1

Relative rates of evolution of noncoding DNA flanking a gene, exons, and introns.

noncoding regions, even though the coding regions are (in general) functionally more important than the noncoding regions.

Thus, when it comes to functional significance, the overall message is that less important proteins/genes (or less important parts of proteins/genes) evolve more rapidly than more important proteins/genes (or the more important parts of proteins/genes). This is precisely the opposite of what would be expected if most of the genetic differences between species are adaptive—such adaptive genetic differences would be expected to have important functional consequences.

CONSERVATIVE SUBSTITUTIONS OCCUR MORE FREQUENTLY THAN DISRUPTIVE ONES

Another prediction we can make is that if genetic differences between species do indeed reflect mostly adaptive substitutions, then we should expect that they would have large effects on protein or gene function. Conversely, if most genetic differences between species don't matter in terms of adaptation or fitness, then we should find higher frequencies of conservative or benign substitutions than those of substitutions with big or disruptive effects. This is not the same as saying that substitutions should occur completely at random, as mutations that interfere with protein function are expected to be selected against under either hypothesis.

The actual evidence overwhelmingly favors the latter prediction, namely, conservative substitutions are far more frequent than disruptive substitutions. As just one example, take a look at Table 6.1, which gives the nonsynonymous substitution rate (i.e., for substitutions that change the amino acid sequence) and the synonymous substitution rate (i.e., for substitutions that do not change the amino acid sequence), based on comparisons between humans and rodents for various proteins. There are two key observations to make from this table. First, for every protein, the synonymous rate is much higher than the nonsynonymous rate. Clearly, substitutions that do not change the amino acid sequence of the protein are much more likely to be conservative than substitutions that do change the amino acid sequence, so the higher rate of synonymous substitutions is consistent with the predictions of the neutral theory. The second key observation from Table 6.1 is that the nonsynonymous substitution rate varies much more between proteins than does the synonymous substitution rate. The nonsynonymous rate varies over 100-fold between proteins (from 0.01 to 1.41), while the synonymous rate varies by less than a factor of 2 (from 3.53 to 5.14). Nonsynonymous rates are roughly correlated with the overall functional importance of the protein, whereas the synonymous rates are roughly constant across proteins, at least when comparing humans to rodents. This remarkable observation leads us to the next main feature of neutral theory, discussed in the next section.

TABLE 6.1 ■ Substitution rates based on comparisons between human and rodent genes

Protein	Nonsynonymous rate	Synonymous rate
Actin α	0.01	3.68
Insulin	0.13	4.02
Growth hormone	1.23	4.95
α-Globin	0.55	5.14
Myoglobin	0.56	4.44
Lactate dehydrogenase A	0.20	5.03
α1 Interferon	1.41	3.53
Apolipoprotein E	0.98	4.04

Source: From Li, W.-H., and Graur, D., *Fundamentals of Molecular Evolution*, Sinauer Associates: Sunderland, MA, 1991.

THE RATE OF MOLECULAR EVOLUTION IS APPROXIMATELY CONSTANT

Suppose we take the amino acid sequences for a particular protein and see how different they are for various pairs of species for which we have some idea of their divergence time, which we can get from fossil or biogeographic evidence. You might think that nothing interesting could come from such comparisons—after all, according to evolution by natural selection, the amino acid sequence of a protein in a particular species should depend on how selection has operated on that protein in that species, so there is no reason to suspect that the number of amino acid differences between two species would have anything to do with their divergence time. And yet, when Emil Zuckerkandl and Linus Pauling did just this in 1962 (Zuckerkandl and Pauling 1962), they found an astonishingly close relationship between the number of amino acid differences in the α-globin and β-globin polypeptides of hemoglobin and divergence time (Figure 6.2). This was the first demonstration that the rate of molecular evolution is constant over time, which in turn leads to the concept of a **molecular clock**: with a molecular clock, we can estimate when two species diverged from the number of differences in their molecular sequences (either amino acid sequences or DNA sequences), even without any fossil evidence. For example, suppose we estimate 75 substitutions between two species; then according to the plot in Figure 6.2, these two species would have diverged about 100 million years ago. This illustrates the power of dating with the molecular clock: in the absence of any relevant fossil evidence, one can nonetheless obtain estimates of species (and, as we shall see in Chapter 12, population) divergence times, and thereby gain insights into their evolutionary history.

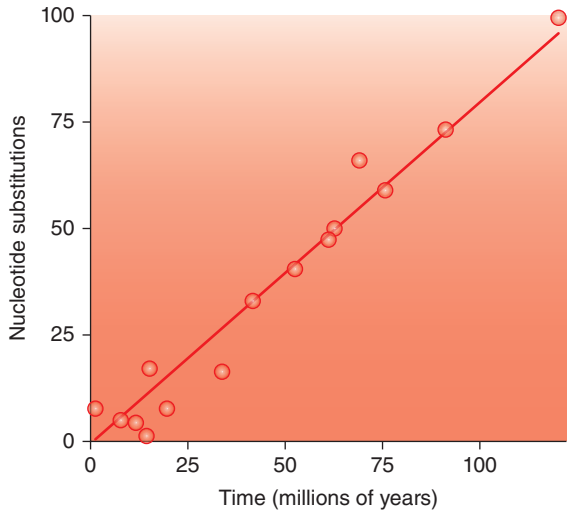


FIGURE 6.2

Plot of the estimated number of nucleotide substitutions versus the divergence time for various pairs of species (points). The relatively good fit to the line indicates that the rate of molecular evolution is roughly constant over time. Modified with permission from Kumar S., "Molecular clocks: four decades of evolution," *Nature Reviews Genetics* 6:654, 2005.

And, while there is no reason to expect that the rate of molecular evolution would be constant over time if most substitutions were adaptive, it is easy to show that a constant rate of molecular evolution is expected if the rate of occurrence of mutations that do not influence fitness is also constant over time. The rate of molecular evolution (i.e., the rate of substitutions in DNA sequences between two species) is the product of two factors: the rate at which mutations occur, and the rate at which new mutations rise in frequency to become fixed in the species. Although we already went through this in Chapter 5, it is worth repeating here (for those who skipped that chapter!): suppose we have a population with effective size N, and neutral mutation rate μ (by neutral mutation rate, we mean the rate of new mutations that are neither advantageous nor disadvantageous to have). Then,

Number of alleles in the population = $2N$
Number of new mutations per generation = $2N\mu$

Probability that a new mutation will reach fixation = $1/2N$ (recall that the probability of fixation of an allele via genetic drift is equal to the frequency of the allele, and by definition a new mutation is present in just one copy in the gene pool).

Rate of molecular evolution = (rate at which new mutations arise) (probability of fixation) = $(2N\mu)(1/2N) = \mu$.

It, therefore, follows that if the neutral mutation rate is constant over time, then the rate of molecular evolution will also be constant over time. It may seem somewhat counterintuitive that the rate of (neutral) molecular evolution does not depend at all on the population size. The reason is that in a small population, there are fewer new mutations occurring each generation but fixation goes more quickly. Conversely, in a big population, there are more new mutations occurring each generation but fixation takes longer. Remarkably, these two processes balance each other exactly, so the overall rate of molecular evolution is the same regardless of the population size.

The molecular clock has been an extremely powerful tool and has provided some important insights into our evolutionary history. For example, as we shall see in Chapter 13, molecular clock approaches

provided the first evidence of a close evolutionary relationship between humans and chimpanzees, as well as strongly supporting a recent African origin of our species. Nevertheless, there are important issues—and limitations—that arise with dating via molecular clocks, and the use—and misuse—of molecular clocks will be discussed in more detail later. For now, just be aware that molecular clocks provide a powerful and important alternative to fossil or archaeological evidence for dating species or population divergence times.

CONTRASTING PHENOTYPIC AND MOLECULAR EVOLUTION

The above sections have laid out evidence to suggest that at the molecular level, most genetic differences between species do not seem to be adaptive. And yet, adaptive evolution clearly occurs. How can we reconcile the Darwinian view of evolution via natural selection on phenotypes, with evolution at the molecular level? To make the problem more concrete, suppose I were to give you several photos of chimps and several photos of humans and ask you to tell me which photos are of chimps and which are of humans. This is something that the average 3-year-old could do, right? And yet, if I were to give you the chimp and human amino acid sequences of a protein, on average 25% of the time you would not be able to tell them apart, because 25% of our proteins have the same amino acid sequence in chimps. At the DNA level, humans and chimps are about 98.4% identical. The (apparently) large phenotypic differences between humans and chimps are thus not reflected in their molecular differences. And this contrast between phenotypic differences and molecular differences is not just because humans pay great attention to even subtle phenotypic differences. Consider Figure 6.3, which compares two frog species to one another and humans and chimps to one another. Clearly, the phenotypic differences between chimps and humans are much bigger than those between the two frog species; yet, the molecular differences are comparable between both pairs of species.

So, phenotypic evolution is not reflected in molecular evolution, and as we have already seen, most molecular differences between species do not appear to influence fitness. Does this mean that classic Darwinian evolution via natural selection—"survival of the fittest"—is wrong? Hardly! For although the bulk of the genetic differences between species (and between populations) conforms to the predictions of neutral theory, there are also exceptions—and lots of them. In fact, one of the major uses (and advantages) of neutral theory is that it provides an easily tested set of predictions concerning various properties of genetic variation within populations/species and genetic differences among populations/species. Looking for genes or mutations that depart significantly from these predictions is an extremely useful way to identify genes or mutations that have been (potentially) influenced by natural selection. For example, it was stated previously that functionally important parts of molecules tend to evolve more slowly than less important parts. However, in certain genes involved in disease resistance, the functionally important parts evolve the fastest, which indeed reflects selection on these genes for disease resistance. Detecting natural selection at the genetic level is an important enough topic that it deserves not just one but two separate chapters and hence will be covered in Chapters 17 and 18. For now, just be aware that there are indeed genetic differences between species and populations that have been influenced by natural selection, and we will see some examples later on.

So how can we reconcile neutral theory with Darwinian evolution via natural selection? Neutral theory is best viewed as an important extension of Darwinian evolution. Most mutations are neutral (or close enough to neutral in their fitness effects that for all practical purposes they are neutral), and hence their fate in populations is governed by genetic drift, migration, and so forth. This is very useful for molecular anthropologists, because it means that the fate of such mutations is influenced by **demography**, that is, population divergence, population size changes, and migration. We can, therefore, use such neutral variation to learn about the demographic history of populations and species. But occasionally, mutations will arise that do have important fitness effects, and because these can be distinguished (at least sometimes) from neutral mutations, by studying molecular genetic variation, we can also learn something about how natural selection has influenced our species. Thus, our genes carry information about both the demographic history of our species and the influence of natural selection, and it is this combination of demography and selection—as revealed in the record of neutral mutations and mutations with fitness consequences, respectively—that comprises our evolutionary past. And as we shall see in the chapters that follow, understanding both our demographic history and the impact of selection is a major goal of molecular anthropology.

And how can we understand the molecular basis of phenotypic evolution, such as phenotypic differences between humans and chimps? We will come back to this issue, but first we need to consider how new gene functions arise.

FIGURE 6.3

Different rates of morphological evolution for pairs of species with approximately the same divergence time, based on DNA sequences. Top left shows a northern leopard frog while the top right shows a southern leopard frog, which diverged at roughly the same time (or perhaps even a little earlier) than the human and chimpanzee shown on the bottom left and right. However, the two frog species are much more similar to one another morphologically than the human and chimpanzee are to one another, indicating much more rapid morphological evolution between humans and chimpanzees than between these two frog species. Top left, top right, and bottom right, reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Lithobates_pipiens.jpg; https://commons.wikimedia.org/wiki/File:Southern_Leopard_Frog,_Missouri_Ozarks.JPG; https://commons.wikimedia.org/wiki/File:Schimpanse_Zoo_Leipzig.jpg); bottom left, photo of the author by B. Pakendorf.

HOW DO NEW GENE FUNCTIONS ARISE?

Another important aspect of molecular evolution is concerned with how new gene functions arise. Clearly new capabilities and features have arisen throughout the evolutionary history of life on this planet, but if an organism's set of genes are already all busy carrying out particular functions, how can genes take on new functions? Multitasking is of course one possibility, and indeed there are genes that carry out multiple functions. But by far the most important mechanism for generating genes with new functions is **gene duplication**. The idea is quite simple: if you have a gene that is constrained because it is already carrying out a

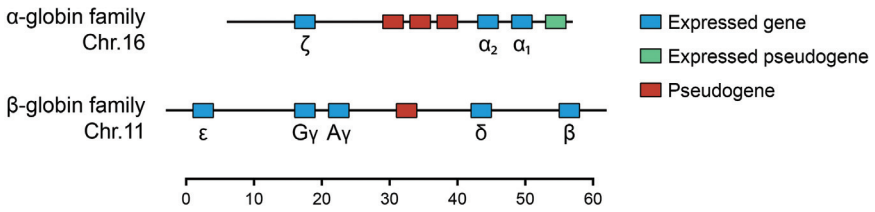


FIGURE 6.4

α-Globin and β-globin gene families, showing the genes and pseudogenes in each family. The scale at the bottom of the figure is in kilobases.

particular function, if you then have an extra copy of that gene, that copy is free to evolve and take on a new function without compromising the original function.

There are a number of ways in which genes can become duplicated, too many to go into detail here. More importantly for our purposes, there is abundant evidence for the importance of gene duplication during evolution. Practically all genes exist as members of **gene families**, which are groups of genes that are structurally (and often, functionally) related. Consider, for example, hemoglobin, which as mentioned previously is responsible for transporting oxygen throughout the body. Hemoglobin consists of four polypeptide chains: two α chains and two β chains. As it turns out, there is not just one gene for the α polypeptide and one for the β polypeptide, but rather an entire family of α genes and another of β genes (Figure 6.4): in humans, there are seven α genes and six β genes, and comparing the exon–intron structure as well as the DNA sequences makes it quite clear that the genes in each family are related to one another by gene duplications. What all these genes are actually doing will be discussed in the following text.

Moreover, the α-globin and β-globin gene families are also related to one another, again via gene duplication. In fact, one can trace the evolutionary history of the globin gene families to find out where the duplication events occurred (Figure 6.5). It turns out that there was a single globin gene that duplicated about 600–800 million years ago to form the ancestral hemoglobin and myoglobin genes; myoglobin is responsible for storing oxygen in muscle tissue and exists as just a single gene in humans. Jawless fish, such as lampreys, have only a single type of hemoglobin, whereas all other vertebrates have at least one α-globin and one β-globin gene, and the duplication event giving rise to the α-globin and β-globin genes occurred about 500 million years ago. Further duplication events took place to produce each family of globin genes; the estimated dates of these duplication events come both from molecular clock dating based on sequence divergence and from the phylogenetic distribution of particular genes—for example, the α_1 and α_2 genes are present in all apes, and hence the duplication must have occurred before the divergence of existing ape species, some 15 million years ago.

And what are all these α-globin and β-globin genes doing? It turns out that some of them are doing nothing at all—they are **pseudogenes**, copies of functional genes that have become nonfunctional because they have accumulated stop codons or other mutations that interfere with gene expression. Our genomes are littered with these nonfunctional copies of genes: there are at least 12,000 pseudogenes in the human genome or almost one pseudogene for each functional gene. Pseudogenes can be thought of as failed experiments in evolution and in all likelihood reflect the usual outcome of gene duplications: throw new mutations at random at a gene, and you are far more likely to break it than you are to come up with something new. But every so often a new mutation does result in something new at the phenotypic level, and once that happens, then natural selection can operate to further develop and refine this new function.

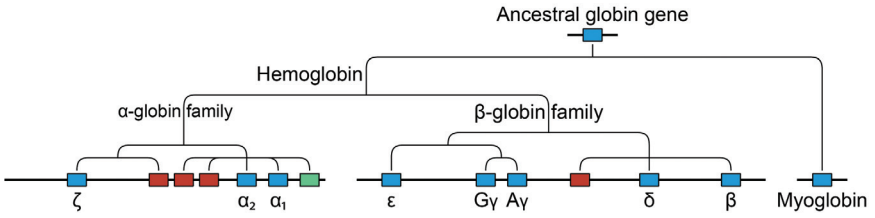


FIGURE 6.5

Evolutionary history and gene duplication events that gave rise to the α-globin and β-globin gene families. Genes and pseudogenes are colored as in Figure 6.4.

The human α-globin and β-globin gene families reflect both processes: of the seven α-globin genes, four are functional and three are pseudogenes, while of the six β-globin genes, five are functional (although one, δ globin, is barely transcribed in humans and is a pseudogene in Old World monkeys) and one is a pseudogene. Moreover, the functional globin genes differ in both when they are expressed during development as well as in their physiological properties. For example, the two gamma genes of the β-globin gene family are expressed during fetal development only—expression stops shortly after birth—and this fetal hemoglobin has a higher affinity for binding oxygen than the adult form of hemoglobin. This makes sense, because the uterine environment is lower in oxygen than outside the womb. Presumably, after a gene duplication event, new mutation(s) arose that increased the hemoglobin oxygen-binding affinity, which in turn was selectively favored in the fetal environment, which then led to further mutations that limited the expression of this new globin gene to the fetus. Other globin genes are expressed only in the embryo, which also has different oxygen requirements. Another example is provided by the ancestral duplication event that gave rise to α-globin and β-globin, as hemoglobin made up of α-globin and β-globin exhibits regulatory capabilities (such as the ability to change how much oxygen is bound according to varying blood chemistry) that are not exhibited by the hemoglobin of jawless fish, which have just one type of globin gene. The take-home lesson: gene duplication allows novel gene functions to arise and be selected for, without disrupting the ongoing functions that the existing genes have to carry out.

Not only do practically all genes occur as members of gene families, but other DNA elements also exist as families related by duplication events. Two important examples are **LINEs** (**long interspersed elements**) and **SINEs** (**short interspersed elements**). These are DNA sequences that are, as the names suggest, either long (several thousand nucleotides in length) or short (a few hundred nucleotides in length) and exist in multiple copies throughout the genome. While several different families of LINEs and SINEs are known that vary considerably in their details, as an illustrative example let's consider a family of SINEs known as **Alu elements**. Alu elements get their name because they typically contain a characteristic DNA sequence that gets cut by an enzyme called AluI (enzymes that cut DNA are called restriction enzymes, and we'll learn more about them in the next chapter). You have more than 1 million Alu elements scattered throughout your genome, each of which is about 300 nucleotides long, which means that about 10% of your genome is just Alu elements. The vast majority of these are inert, apparently doing nothing more than just taking up space in your genome. However, a few of these are

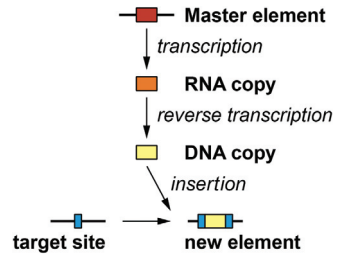


FIGURE 6.6

Retrotransposition: a master element is transcribed to make an RNA copy, which is then reverse transcribed to make a DNA copy, which then inserts into a new site (the target site) in the genome, thereby creating a new element.

active, in the sense that every now and then they make a new Alu element by a process known as **retrotransposition** (Figure 6.6): first an RNA molecule is transcribed from one of these "master" Alu elements, then DNA is synthesized from the RNA molecule via a process known as reverse transcription (which, you will recall from Chapter 2, goes against the Central Dogma!), then this new Alu element goes off and inserts into a new location in the genome (these "daughter" elements are not capable of further transposition; only the master elements have the additional functional sequence elements necessary for retrotransposition to occur). So, over evolutionary time Alu elements have been increasing in number, although occasionally Alu elements are deleted. Alu elements are found in all primates (but no nonprimates), so this process of accumulation of Alu elements has been going on since the origin of primates, some 80 million years or so ago.

An interesting—and as yet, unanswered—question is whether or not Alu elements (and other transposable elements) provide any useful functions. For sure they can be detrimental, as examples are known of genetic diseases in children caused by a novel insertion of an Alu element into a gene that disrupts its function. One school of thought holds that Alu elements are purely "selfish" elements; they don't do anything useful but exist because the master elements are continually pumping out new copies. We would be better off if we could just delete the master elements, but new master elements arise faster than our genome can get rid of them. However, another school of thought holds that Alu elements actually play some important roles. For example, Alu elements promote genomic rearrangements, which can lead to large-scale effects in gene regulation—move a gene into a new chromosomal environment, and you may very well alter when and where it is expressed, which can have important evolutionary consequences, as discussed in the next section. For now, the jury is still out, but regardless of

the explanation for why Alu elements exist, they are yet another example of the mysteries of our genome.

GENE REGULATION AND PHENOTYPIC EVOLUTION

We've already remarked on the discrepancy between the phenotypic versus molecular differences between chimps and humans. This was actually pointed out in 1975 by Mary-Claire King and Allan Wilson, who noted that the number of amino acid differences in the proteins of chimps and humans seems to be too small to account for their phenotypic differences (King and Wilson 1975). Now, these results were based on a small sample of proteins, so a potential explanation was that other proteins harbored more (or more significant) amino acid differences that accounted for the phenotypic differences between chimps and humans. But King and Wilson made the provocative suggestion that perhaps the important phenotypic differences between chimps and humans were not caused by structural changes (i.e., amino acid differences) in their proteins. Instead, they proposed that differences in gene regulation—how much of a protein was made, and in what tissues, and when during development—could be responsible for the important phenotypic differences between humans and chimps. We have already seen with the globin genes how differences in gene regulation can arise and how important they can be, and it is easy to speculate how such changes could apply to the issue of chimp–human differences. For example, the larger brain of humans compared to chimps mostly reflects differences in growth rates: both human and chimp brains grow at about the same (fast) rate during fetal development, but following birth the rate of growth of chimp brains slows considerably, while human brains continue growing at the same fast fetal rate for the first year of life. At the moment we don't know how this occurs, but an intriguing possibility is that the growth factor proteins are more or less the same for chimps and humans, but humans continue expressing these growth factors through the first year of life, while chimps stop expressing them after birth. Thus, simple changes in gene regulation could, in principle, account for complex phenotypic differences such as the larger brain size of humans. Who knows—perhaps we could take all of the genes that a chimp has, and simply by altering their expression and regulation, end up with a human! We will return to the topic of gene regulation differences during evolution in Chapters 18 and 20. For now, the important take-home message (also made in Chapter 2, but I can't resist repeating it here) is that when it comes to phenotypic evolution—as in much of life—maybe it's not what you have, but rather what you do with what you have, that counts.

CHAPTER 7

GENETIC MARKERS

We've now covered the basics of what genes are, what they do, how they do it, how they behave in populations, and how they evolve. But there is still a way to go before we can discuss the insights that molecular anthropology has provided into human evolution and population history. First, we need to consider the various types of **genetic markers** that have been analyzed in molecular anthropology studies and their properties (which is the subject of this chapter). Next, we need to say a few words about sampling populations and genomic regions (the subject of the next two chapters), as well as the various methods used to analyze and make inferences about population history from molecular genetic data (covered in later chapters). To start off the discussion of genetic markers, we can distinguish between genetic markers based indirectly on variation in the products of genes (also known as **classical markers**) versus genetic markers based directly on variation in DNA.

CLASSICAL MARKERS: IMMUNOGENETIC MARKERS

"Classical marker" is a catchall term to denote any genetic marker based on variation in the products of genes, as opposed to variation at the DNA level. We've already introduced the first such marker to be used in studies of human variation, namely, the ABO blood groups. After the discovery of the ABO blood groups by Karl Landsteiner in 1900, a Polish physician by the name of Ludwik Hirszfeld—who sometimes used the German spelling of his last name, Hirschfeld—showed that they were inherited as Mendelian characters (although, as we saw in Chapter 4, they were first wrongly assumed to be governed by two loci, each with two alleles). Hirszfeld was stationed at a hospital in Serbia during World War I, during which time he pioneered the use of blood transfusions—often using his own blood. Many soldiers and refugees from various nations passed through the hospital at the end of the war, and Hirszfeld realized that there was an unparalleled opportunity to investigate variation in ABO blood group frequencies in different populations. He and his wife Hanka (also a physician) conducted the first study of genetic variation in humans, tabulating the ABO blood group frequencies in more than 8000 people. The results are given in Table 7.1, and there are several noteworthy features: samples sizes are large (numbering in the hundreds); populations are described relatively precisely; and numerous populations are included. By contrast, many subsequent studies have often included much smaller sample sizes from many fewer populations that are much less precisely described (e.g., it is not uncommon to find studies of some genetic markers in a few samples from populations described only as Africans, Europeans, or Asians). Despite this being the seminal study of human genetic variation, the Hirszfelds ran into difficulties in getting their study published. They first sent the manuscript to the premier medical journal of the time, *The British Medical Journal*, only to have the manuscript languish for several months before finally being returned with the editor replying that the results were not of interest to the medical community (a sentiment that alas is still shared by some journal editors when it comes to studies of human genetic variation!). Fortunately, the Hirszfelds persevered and remained convinced that their approach had value; as Ludwik presciently wrote a few years later, "I do not doubt that the blood groups can help to solve the deepest problems of anthropology" (quoted in Allan 1963).

The ABO blood groups are one of more than 30 known blood group systems (the Rh and MN blood group systems are others that we've already seen), all based on various antigens found on the surface of red blood cells. Several of these are variable enough to have been useful in studies of human genetic variation. Blood groups in general are examples of what are known as **immunogenetic markers**, which share in common the characteristic that they are based on the

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

TABLE 7.1 ■ Frequencies of the ABO blood groups in various human populations

	A	B	AB	O	
	In percent	In percent	In percent	In percent	Total no. examined
English …	43.4	7.2	3.0	46.4	500
French …	42.6	11.2	3.0	43.2	500
Italians …	38.0	11.0	3.8	47.2	500
Germans …	43.0	12.0	5.0	40.0	ca. 500
Austrians …	40.0	10.0	8.0	42.0	?
Serbs …	41.8	15.6	4.6	38.0	500
Greeks …	41.6	16.2	4.0	38.2	500
Bulgarians …	40.6	14.2	6.2	39.0	500
Arabs …	32.4	19.0	5.0	43.6	500
Turks …	38.0	18.6	6.6	36.8	500
Russians …	31.2	21.8	6.3	40.7	1000
Jews …	33.0	23.2	5.0	38.8	500
Malagasies	26.2	23.7	4.5	45.5	400
Negroes (Senegal)	22.6	29.2	5.0	43.2	500
Annamese	22.4	28.4	7.2	42.0	500
Indians …	19.0	41.2	8.5	31.3	1000

Source: Reprinted from Hirschfeld, L., and Hirschfeld, H., "Serological differences between the blood of different races. The result of researches on the Macedonian front." *Lancet* 2:675, 1919.

recognition of antigens by antibodies. Other immunogenetic markers that have been used extensively in human genetic variation studies include the antigens originally found on white blood cells, known as **HLA** (human lymphocyte antigens). There are several HLA loci, grouped into different classes related to their properties. Class I loci include the HLA-A, -B, and -C loci, which are found on all cells with nuclei (hence, not on red blood cells, which lack a nucleus) and are involved in defense against viruses. Class II loci include the HLA-DP, -DQ, and -DR loci, and their antigens are found only on cells (principally **lymphocytes**) that are actively involved in recognizing and destroying foreign cells (bacteria and the like). The HLA Class I and Class II loci comprise the human major histocompatibility complex (**MHC**), a family of genes involved in resistance to infectious disease, regulating interactions among different cell types, and rejection of organ transplants—although obviously the MHC did not evolve for this last purpose! Instead, the MHC is involved in distinguishing "self" (i.e., your own cells) from "nonself" (e.g., bacteria and other foreign invaders), and organ transplants unfortunately fall into the latter category. HLA loci are noteworthy for being among the most polymorphic loci known in humans—the HLA-B locus, for example, has more than 2000 described alleles. The extraordinary variability of the HLA loci is generally attributed to heterozygote superiority, in which you will recall that heterozygotes have the highest fitness, although other mechanisms may also play a role.

Another example of immunogenetic markers is antigens found on antibodies themselves (which, confusingly, are also detected by reactions with antibodies!), of which the principle examples are the **Gm** and **Km immunoglobulin allotypes**. The genetic variants in these two systems are inferred from the combination of antigens detected by various antibodies, and there are numerous alleles described at both the Gm and Km loci, which made them useful in studies of human genetic variation.

Through the latter half of the twentieth century, many studies of genetic variation in humans were carried out via immunological typing of blood group, HLA, and immunoglobulin loci. While these studies were important for providing the first insights into the genetic structure of human populations, there are two complications associated with immunogenetic markers. First, they require a **typing serum**—that is, a source of the antibody that is used to detect a particular antigen. There are various ways to do this, but a common way of obtaining antibodies is to inject human cells with the antigen of interest into an animal (mouse, rabbit, goat, etc.), which will then produce antibodies against the antigen, which are then obtained from the blood of the injected animal. Sounds good in principle—but in practice, different animals may respond somewhat differently to the same antigen, so the quality and properties of the typing serum may vary. Plus, you only get so much typing serum from one animal, so eventually you will run out of a typing serum and have to make more,

and the new typing serum might not react in the same way as the old one. So, using different typing sera for supposedly the same antigen can easily produce different results—standardization of results thus becomes an issue. Second, as with any biological assay, there is an element of the unknown concerning what is really happening in the assay. You mix a typing serum containing one or more antibodies with some cells and get a positive reaction—but is the antibody really binding to the antigen of interest, or is there something else in the sample that it is binding to? This is of particular concern when a typing serum obtained by injecting cells from someone of one ancestry (e.g., European) is then used to detect antigens on cells from people of a different ancestry (e.g., African); it has happened in such cases that the antibody reacts positively but to a different antigen, leading to the mistaken conclusion that the same antigen exists in different populations.

For these reasons, and for the added reason that analyzing DNA is now faster, cheaper, and easier than analyzing proteins or antigens, practically all assays of immunogenetic markers are nowadays carried out via genotyping or sequencing the underlying DNA sequence variants. DNA analysis methods will be discussed later in this chapter. Before we leave the topic of immunogenetic markers, though, there is one other interesting aspect worth pointing out. Figure 7.1 shows the structure of a typical antibody and various MHC molecules and receptors. Note that the structures share some superficial similarities that did not arise by chance; as Figure 7.2 shows, the genes that encode antibodies and those that encode MHC molecules are related by gene duplication events. In fact, there exist a tremendous variety of immunogenetic proteins (only some of which are shown in Figure 7.2), all specialized to carry out various tasks in distinguishing self from nonself and getting rid of foreign cells, and all part of one big gene family, related by gene duplication events. The family of immunogenetic proteins (sometimes referred to as a superfamily, since antibodies, MHC molecules, etc., can be thought of as separate multigene families in their own right) thus represents an excellent example of the important role of gene duplication in generating new functions, as discussed previously in Chapter 6.

CLASSICAL MARKERS: BIOCHEMICAL POLYMORPHISMS

In addition to immunogenetic markers, there are other types of variation in the protein products of genes that have been used in studies of human genetic variation. These are generally referred to as **biochemical polymorphisms**—not a particularly informative name as all polymorphic variation in either genes or their products are biochemical polymorphisms, since both DNA and proteins are biochemicals. But the general idea behind the name is that biochemical polymorphisms reflect variation in the amino acid sequence of proteins, which in turn reflect underlying DNA sequence variation. The basic idea is that proteins carry an electric charge, since some of their constituent amino acids also carry electric charges. So, put a protein in an electric field under the right conditions and it will migrate toward either the positive or the negative pole—this procedure is known as **electrophoresis**. And, the rate of migration will depend both on the

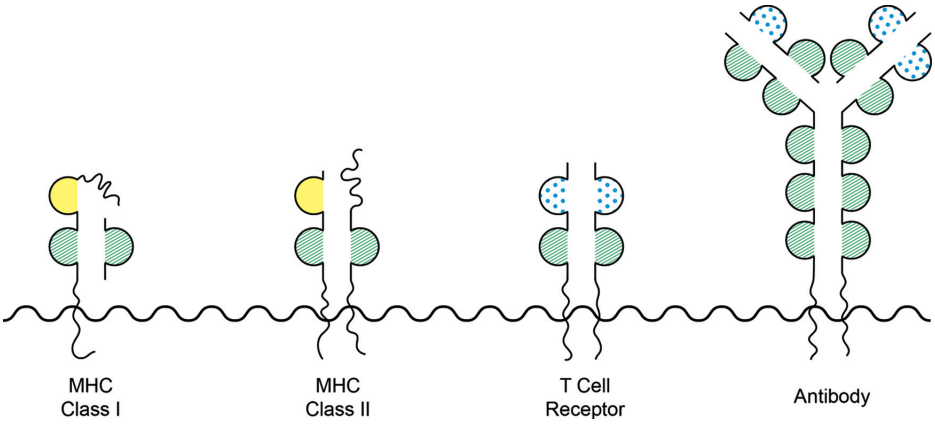


FIGURE 7.1

Schematic structure of various molecules involved in the immune response: MHC class I and class II, T-cell receptor, and antibody molecules, showing similarities in structure. These molecules are all cell surface receptors, meaning that part of the protein is inside the cell and part lies outside the cell. Similarities in parts of the protein structure (domains) are indicated by different colored shading (yellow, green, and blue colors). Different receptors share similar domains, indicating descent from a common ancestral molecule. MHC, major histocompatibility complex.

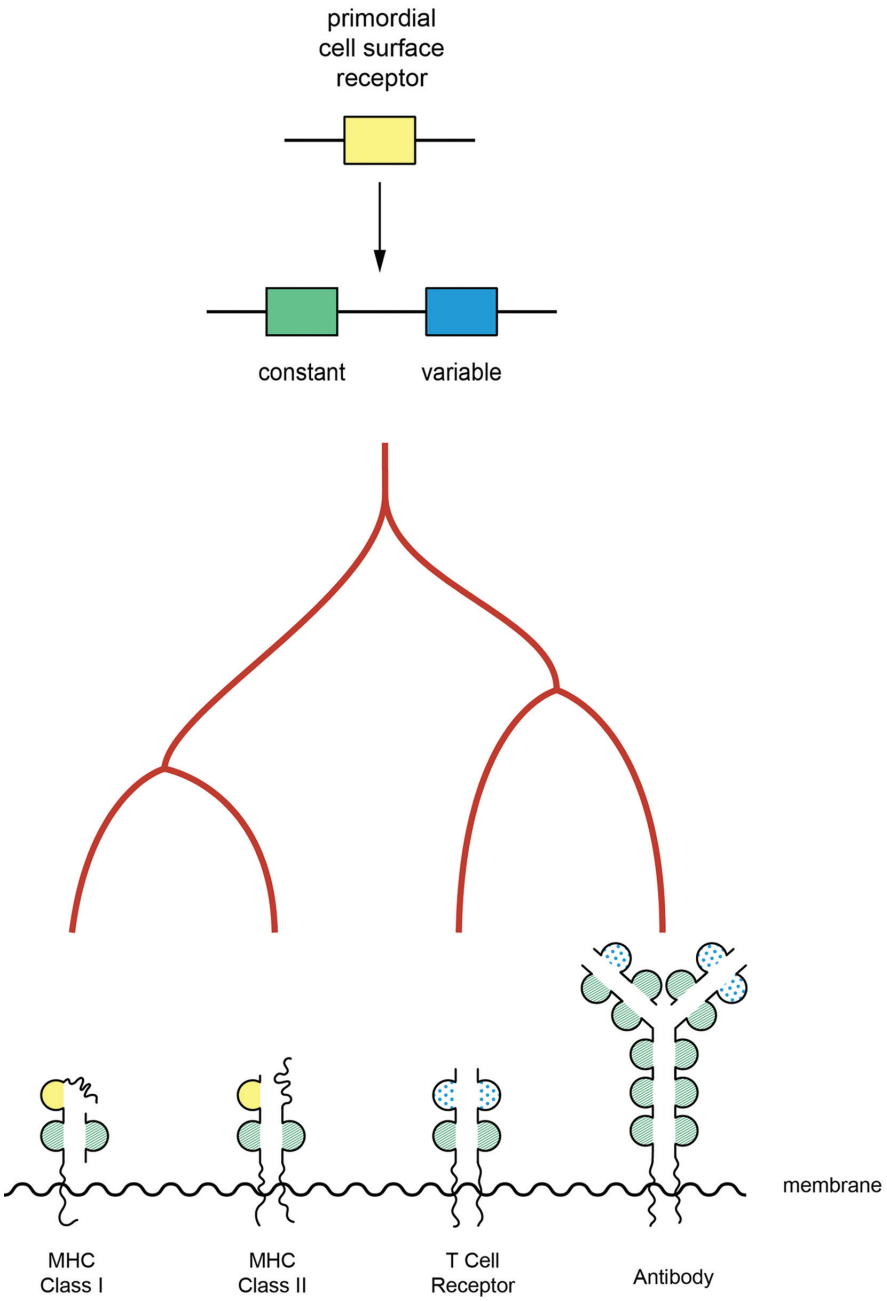


FIGURE 7.2

Evolutionary history of various immunological molecules, illustrating gene duplication events from a primordial cell surface receptor that duplicated to form the different types of domains that are characteristic of these molecules. MHC, major histocompatibility complex.

electric charge of the protein and its overall shape. Suppose now that we have a mutation in the gene for this protein that changes not only the amino acid sequence of the protein but also the overall shape or electric charge of the protein. That protein will now migrate somewhat differently in the electric field, that is, it will have a different electrophoretic mobility. Thus, by measuring changes in the electrophoretic mobility of proteins, we can detect genetic variation (how this is actually done will be discussed shortly).

Two big breakthroughs that occurred in the 1950s made it feasible to screen lots of loci for genetic variation in all sorts of creatures (including humans), thereby revolutionizing our understanding of the prevalence and extent of genetic variation in natural populations. The first vastly simplified the process of electrophoresis by introducing gels made of starch as a fast, reliable, cheap, and easy to use means of carrying out protein electrophoresis. Oliver Smithies (who later won a Nobel Prize for developing methods that

allow specific genes to be modified in cells) came up with this idea (Smithies 1955), based on his recollection of the starch his mother used in doing the laundry when he was a child—great ideas often have humble origins! Other materials that form suitable gels were subsequently developed, such as polyacrylamide and agarose, which are particularly useful for DNA electrophoresis, but starch gels dominated protein electrophoresis studies.

The second big breakthrough had to do with visualizing the proteins after electrophoresis. To measure the electrophoretic mobility of proteins, we need some way to detect them after electrophoresis. One way to do this is to use a dye that binds to any protein. While all proteins will be stained with such a dye, proteins present in higher concentrations will be stained more strongly, and since we already know a lot about which proteins are present at high concentration in blood, it is relatively straightforward to assign various bands to specific proteins. From such **electropherograms**, one can then look for shifts in the electrophoretic mobility of particular proteins in different individuals, and (as described later) use this information to assign genotypes to individuals.

However, the vast majority of proteins are present at too low a level to be detected by such general protein stains, so with this method you typically detect only a few proteins. This limitation was circumvented by the development of **histochemical stains**, which are based on the activity of a specific enzyme. Recall that enzymes are a special type of protein that enable particular chemical reactions. Enzymes typically greatly accelerate the rate at which one biochemical is converted into another biochemical; for example, the enzyme lactate dehydrogenase (LDH) converts pyruvate to lactate, and in the process a free hydrogen ion is generated. So if you add the appropriate chemicals to the starch gel after electrophoresis, and include a soluble dye that takes up the free hydrogen ion generated by LDH activity and thereby becomes insoluble, then—like magic—you will see dark-staining bands in the starch gel corresponding to where the LDH protein has migrated during electrophoresis. The advantages of histochemical staining are that a specific enzyme is targeted (based on the chemical reaction catalyzed by that enzyme), and the amount of staining is proportional to the activity of the enzyme, not how much of the enzyme is present, so even minute amounts of an enzyme can be detected.

It turns out that many enzymes carry out reactions that can be linked to a dye, and there are recipes for histochemical stains for more than 100 enzymes. Applying these various stains to starch gels revealed two types of variation. Some enzymes exist in multiple molecular forms (i.e., with different amino acid sequences), where each form is the product of a different genetic locus—these different forms are called **isozymes**. For example, in humans there are three loci for LDH, and they are predominantly (albeit not exclusively) expressed in different tissues: LDH-A has the highest expression in muscle tissue, LDH-B has the highest expression in heart tissue, and LHD-C is expressed in the testis (other tissues do express these loci to varying degrees). And—as you might suspect by now—these forms are related by gene duplication events, so the LDH isozymes are yet another example of a multigene family, with different loci adapted for expression in different tissues.

The second type of variation revealed by histochemical staining of enzymes after starch gel electrophoresis is the type alluded to at the beginning of this section on biochemical polymorphisms, namely, mutations at a single genetic locus that result in amino acid substitutions that alter the electrophoretic mobility of the protein. Different forms of a protein that reflect alleles at a single genetic locus are called **allozymes**. How can we use allozyme variation to infer genotypes? Consider the (hypothetical) example depicted in Figure 7.3, which diagrams the sorts of results one might see after staining a starch gel for a particular enzyme. Each individual has one or two bands of enzyme activity, and altogether there are three different zones of activity that can be distinguished (labeled a, b, and c in Figure 7.3). The interpretation is straightforward: this enzyme is encoded by a locus with three alleles that can be distinguished by their electrophoretic mobility, so homozygotes for an allele will have one band of activity, while heterozygotes will have two bands. If we label the alleles that correspond to bands a, b, and c as A, B, and C, respectively, then an individual with just the a band has the AA genotype, an individual with both a and b bands has the AB genotype, and so forth. Electrophoretic mobility variants thus behave

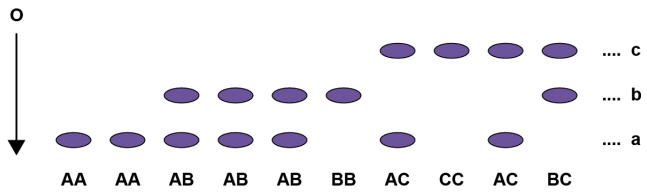


FIGURE 7.3

Banding pattern obtained after histochemical staining for an enzyme with three alleles. The "O" indicates the origin, that is, where the samples were inserted into the gel, and the arrow indicates the direction of migration of proteins. There are three electrophoretically distinguishable forms for this protein, designated a, b, and c (in order of the migration distance from the origin, with a migrating the farthest). Each lane is the result for an individual, with diploid genotypes immediately obvious from the banding pattern (homozygotes have one band, heterozygotes have two bands).

as Mendelian codominant traits, which is quite valuable because an individual's genotype can be readily determined just by inspection of the electropherogram (contrast this with ABO blood groups, for example, where a person with type A blood can be either an AA homozygote or an AO heterozygote).

It should be pointed out that for some enzymes the banding pattern can be much more complex—additional bands can appear if the active enzyme is made up of multiple subunits, or if further posttranslation modifications of the enzyme occur. But any doubts as to how to interpret banding patterns in terms of genotypes can usually be resolved easily with family data. It should also be pointed out that electrophoresis will detect only a fraction of the genetic variation at a locus. Mutations that do not alter the amino acid sequence of the protein, or that result in amino acid substitutions that do not alter the electrophoretic mobility of the protein, will not be detected.

Nonetheless, gel electrophoresis combined with histochemical staining ushered in a new era in studies of genetic variation and not just in humans. It became possible to study genetic variation in any organism that could be ground up (or whose tissues could be ground up) and the proteins extracted and subject to electrophoresis. Beginning with seminal studies in 1966 of genetic variation in fruit flies by Richard Lewontin and Jack Hubby (Lewontin and Hubby 1966), and in humans by Harry Harris (Harris 1966), the ensuing decade saw a flood of papers on the theme of "Genetic Variation in _____" (fill in the blank with your favorite organism). And what these studies invariably found was lots and lots of genetic variation, much more than had previously been suspected by most geneticists. This in turn ushered in a long-standing debate as to whether all of this genetic variation was maintained by natural selection, or if it was largely neutral. As we saw in the previous chapter on molecular evolution, this debate has been resolved, as most genetic variation fits the predictions of neutral theory—although, as we shall see in Chapter 19, there are still contrasting views over the extent to which selection has influenced, and might be continuing to influence, human evolution.

The study of classical markers has been eclipsed almost entirely by studies of variation in DNA, which we will turn to next. Nevertheless, classical markers were a very important chapter in studies of human genetic variation, and numerous insights and hypotheses concerning human population relationships and migrations arose from such studies. The serious student of molecular anthropology would be well-served by perusing an excellent compilation and synthesis of what we learned from classical markers entitled *The History and Geography of Human Genes*, by Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. Much (but by no means all) of what we have subsequently learned from DNA variation is anticipated by this book. And for the not-so-serious student, there is the highly readable and entertaining *Genes, Peoples and Languages*, also by Luca Cavalli-Sforza, one of the pioneers and leaders in genetic approaches to studying human population history.

THE FIRST DNA MARKERS: RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

At about the same time that protein electrophoresis studies were gaining momentum, new developments were enabling the manipulation and study of DNA. A key event was the discovery in 1970 of **restriction enzymes** in bacteria (Smith and Wilcox 1970). Restriction enzymes are proteins that recognize (bind to) a specific sequence in DNA and restrict (cut) the DNA at (or near) that sequence. For example, the restriction enzyme EcoR1 recognizes the sequence GAATTC and will cut a double-stranded DNA molecule wherever this sequence occurs. Note that the sequence is palindromic, in that the DNA sequence on the complementary strand is CTTAAG, which when reversed (so it has the same directionality as the other strand) is also GAATTC. Moreover, as shown in Figure 7.4, when EcoR1 cuts the DNA molecule, it does so in such a way as to leave overhanging ends. These overhanging or "sticky" ends are complementary to, and hence can base-pair with, other DNA fragments produced by digesting DNA with EcoR1. This is the basis of "recombinant DNA," or genetic engineering: take a plasmid (a small, circular DNA molecule from bacteria that typically contains a gene for resistance

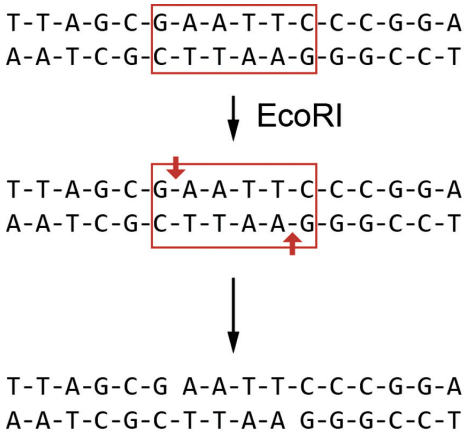


FIGURE 7.4

Digestion of DNA with the restriction enzyme EcoR1. The sequence GAATTC is cut by EcoR1 at the places indicated by the red arrows, producing the fragments at the bottom. Note the palindromic nature of the sequence cut by EcoR1: the sequence on the bottom strand is simply the reverse complement of the sequence on the top strand.

to an antibiotic—tetracycline, for example) that has a single EcoR1 restriction site, so digesting the plasmid with EcoR1 results in a single linear piece of DNA. Add to this DNA from another organism that has been similarly digested with Eco R1, along with the appropriate enzymes and chemicals so that DNA synthesis can occur, and you can get recombinant molecules that consist of the plasmid DNA plus DNA from another organism. Transfer the plasmids to bacteria and grow the bacteria in the presence of tetracycline, and only those bacteria that contain the plasmid can grow. Isolate the plasmids and test them, and some of them will contain the foreign DNA—you now have bacteria that contain something that never existed before in nature, namely, DNA from two different organisms. And if you use the right kind of plasmid and insert DNA that encodes a protein, such as the human insulin gene, you can induce the bacteria to start making human insulin.

Our interest in restriction enzymes stems not from their use in making recombinant DNA, however, but rather in how they can be used to detect DNA polymorphisms. Take DNA from a human and digest it with EcoR1, and you will get a large collection of fragments of various sizes—roughly, 800,000 fragments with an average size of about 4000 base pairs (4 kilobases, or 4 kb for short). If it is not obvious where these numbers come from, consider that the probability of finding the EcoR1 recognition sequence, GAATTC, is $^1/_4 \times {}^1/_4 \times {}^1/_4 \times {}^1/_4 \times {}^1/_4 \times {}^1/_4 = (0.25)^6$, because there is a $^1/_4$ chance of having a G at a position in a DNA sequence, a $^1/_4 \times {}^1/_4$ chance of having G followed by A, and so forth; multiply this average probability of finding GAATTC in a DNA sequence by the number of DNA bases in the human genome, namely, 3.2 billion, to get the numbers above. It turns out that DNA also carries an electric charge, so if you put the digested DNA into an appropriate gel (agarose being the gel of choice for DNA) and subject it to an electric charge, the DNA will migrate, with smaller fragments migrating faster than larger fragments. So, with DNA electrophoresis we get separation of the different fragments, but as with protein electrophoresis there is still the problem of how to visualize the molecules. The solution to this problem for DNA is to use a **probe**, namely, a segment of purified DNA that is made radioactive by the incorporation of radioactive nucleotides. As shown in Figure 7.5, after electrophoresis is completed, the DNA

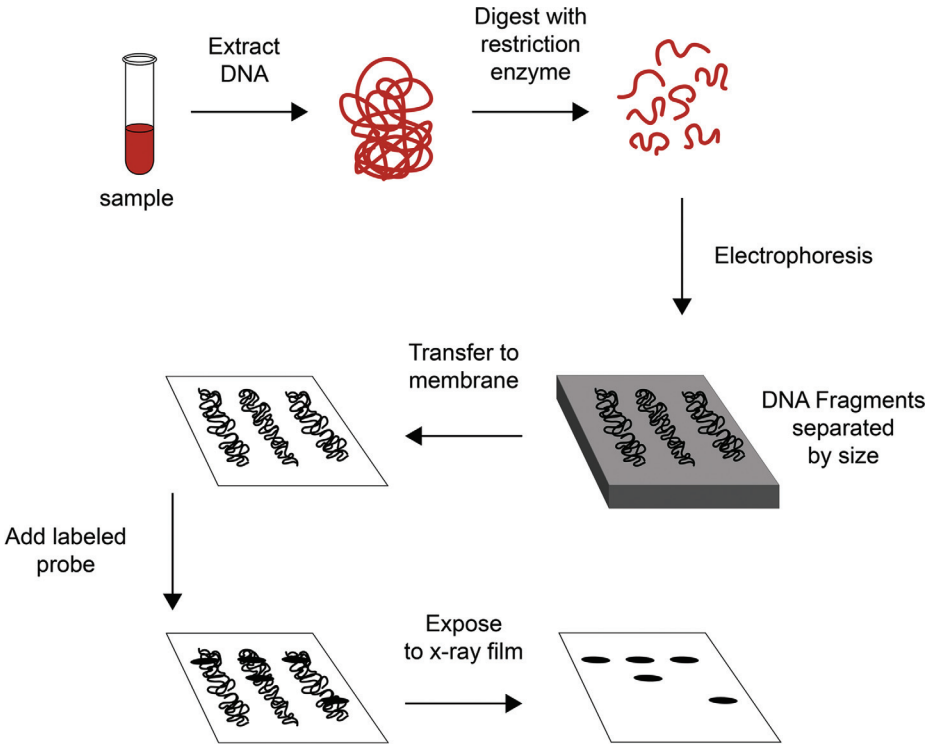


FIGURE 7.5

Steps involved in the Southern blot procedure for visualizing the products of DNA digestion with a restriction enzyme. After purification of DNA and digestion with a restriction enzyme (such as EcoR1, see Figure 7.4), gel electrophoresis is used to separate the DNA fragments by size. The fragments are then transferred to a membrane ("blotted"), which is then incubated with a radioactive probe DNA that detects a specific DNA sequence by binding to it via base-pair complementarity. The radioactive probe, and the DNA fragments it thereby binds to, can be visualized after exposing the blot to X-ray film. See also Figure 7.6.

in the gel is first denatured—made single-stranded—and then transferred by capillary action to a special paper or membrane made of nitrocellulose or nylon—basically, the membrane is used to "blot" the gel. The membrane, which has the DNA immobilized, thereby preserving the size separation of the DNA fragments, is then treated with the radioactive, single-stranded probe under appropriate conditions so that the probe will hybridize to wherever there is complementary DNA on the membrane. Wash off excess probe, expose the membrane to X-ray film, and—like magic!—dark bands will appear on the X-ray film. The inclusion of molecular weight size markers (typically, fragments of viral DNA of known size that are also radioactive) allows the estimation of the size of the DNA fragments. And if you do this for a collection of individuals, you will find differences in the bands that are present—sometimes a band will be missing, sometimes there might be an extra band or two. There are various changes that can cause these differences: mutations that create or destroy a recognition sequence for the restriction enzyme used, insertion or deletion of DNA in the region detected by the probe, and so forth. But what is important for our purposes is that these banding patterns are inherited as simple Mendelian, codominant markers, and hence genotypes can be readily assigned to the banding patterns (Figure 7.6). These polymorphisms have the unwieldy name of "restriction fragment length polymorphisms," or RFLPs (pronounced "riflips" by the cognoscenti) for short.

The procedure described previously for detecting DNA RFLPs is referred to as Southern blotting, named for Ed Southern, who invented the method in the mid-1970s (Southern 1975); Southern also invented another technology for detecting DNA variation, microarrays, described later in this chapter. As a play on Southern's name, when a method was later developed for RNA electrophoresis and blotting, it was called Northern blotting, and a similar method for analyzing proteins via blotting is known as Western blotting (and who says scientists don't have a sense of humor!). Anyway, Southern blots became a standard methodology in molecular biology laboratories for many years, with all sorts of applications. The first genetic markers used for clinical diagnosis of particular diseases, as well as the first genetic markers used for forensic casework, were RFLPs, and numerous genetic variants were discovered and analyzed in humans and other organisms via Southern blots. Still, Southern blots were cumbersome and time-consuming to produce, not readily automated, and required handling radioactive chemicals, something which most of us would prefer to avoid if possible. It therefore should come as no surprise that Southern blots were readily eclipsed by the next technological advance, namely,

FIGURE 7.6

Results of a Southern blot analysis. The numbers on the left indicate the sizes of DNA fragments, determined by a molecular weight size marker (i.e., DNA fragments of known size) that was electrophoresed along with the DNA samples of interest. "Origin" indicates where the digested DNA was inserted into the gel; after applying an electric current, small DNA fragments move through the gel more rapidly than large fragments, thereby separating DNA fragments by size. Lane 1 illustrates an individual homozygous for the presence of a restriction site that creates two fragments of 7.6 and 5.4 kb. Loss of this restriction site results in a fragment of 13.0 kb; Lane 2 is from an individual who is heterozygous (one chromosome has the site, one chromosome lacks it) and so produces the 7.6 and 5.4 kb fragments from the chromosome with the site and the 13.0 kb fragment from the chromosome without the site. Lane 3 is from an individual who is homozygous for the absence of the site and hence has only the 13.0 kb fragment. The other bands in this Southern blot at 14.5 and 4.2 kb are from flanking DNA fragments that are included in the probe sequence but not influenced by variation at this restriction site.

the invention of the polymerase chain reaction (PCR), which we turn to now.

POLYMERASE CHAIN REACTION

Suppose we have a particular target region of the human DNA genome that is a few hundred to a few thousand base pairs long, and we wish to investigate this target region for DNA variants. If we know the DNA sequences that flank our target region, we can take **oligonucleotide primers** (short synthetic pieces of single-stranded DNA that are easy to make and cheap to buy) that are complementary to the flanking DNA and will direct DNA synthesis (i.e., construct a complementary strand of DNA, as described in Chapter 2) across the target region (see Figure 7.7). Since

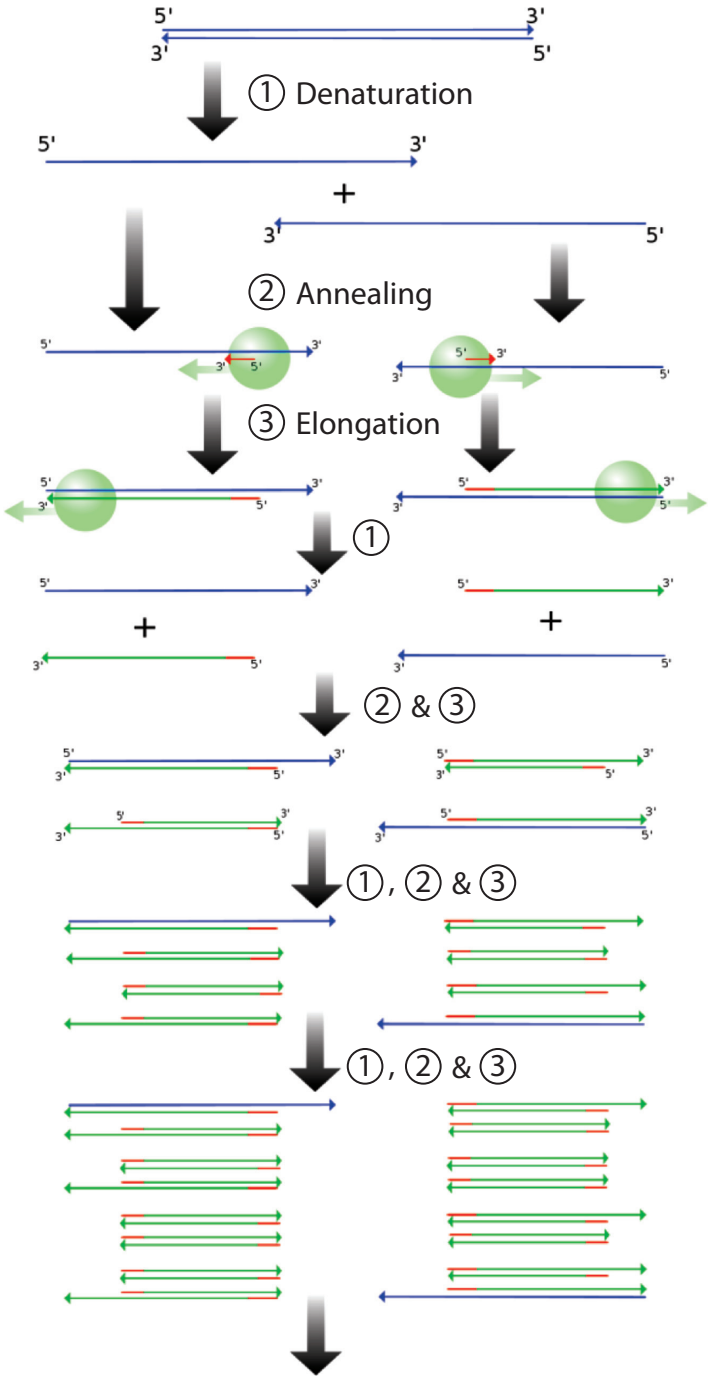


FIGURE 7.7

Steps involved in carrying out the polymerase chain reaction (PCR). In the first step, DNA is denatured (separated into single strands) by heating, typically to 95° C. In the second step, the DNA is cooled, allowing primers (red lines in the diagram) to anneal to the DNA. These primers are short synthetic pieces of single-stranded DNA, typically around 25–35 nucleotides long, that are complementary to the DNA sequences that flank the desired target region. In the third step, DNA polymerase (green circles) replicates the DNA starting from the primers, resulting in elongation of DNA across the target region. This completes the first cycle of the PCR; these steps are then repeated. In the second cycle, primers will anneal again to the target DNA, as well as to the DNA templates that were created in the first cycle. Elongation of the latter will create DNA molecules with defined ends corresponding to the primer sequences and containing the desired target DNA. In subsequent cycles, the number of such DNA molecules will double each cycle, if the efficiency of the process is 100% (which it never is, but after the typical 30–40 cycles of PCR, there will be such a huge increase in the desired PCR product that for all practical purposes the result is a pure DNA preparation of the target DNA). Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:PCR.svg).

DNA strands have directionality and DNA synthesis will proceed only in one direction, this is trivial to arrange. Take our genomic DNA sample that contains our target region (along with the rest of the genome), add primers, DNA polymerase, and nucleotides, heat it up to denature the DNA (i.e., break the bonds that hold the two DNA strands together, thereby making it single-stranded, easily accomplished by heating the DNA to 95°C for a minute or so), and then let the mixture cool to allow the primers to anneal to their complementary sequence and DNA synthesis to occur. After this first cycle, there will be new DNA strands that start with each primer and extend through the target region (see Figure 7.7). Now, heat again to denature the DNA and cool again to allow more primers to anneal and DNA synthesis to occur; in addition to primers annealing to the genomic DNA as before, primers will also anneal to the DNA synthesized in

the first cycle (again, see Figure 7.7). After this second cycle, there will be new DNA strands whose ends correspond to the primers; hence our target region will be contained in a DNA fragment of a defined size—the PCR product, or **amplicon**. Continue this process of heating and cooling, and each subsequent cycle will (theoretically) double the amount of PCR product. After 20 cycles, for each starting (template) DNA sequence, there would be up to 1 million copies of the amplicon containing the target region; after 30 cycles, there would be more than 1 billion copies. Essentially, what one has after PCR is a sample that contains so many copies of the target DNA of interest that the remaining DNA in the sample is negligible. It's like taking a page of this book and making a billion photocopies of it and sticking them all into this book—for all practical purposes, the book would then consist of only that one page.

Now that we have a sample with billions and billions of copies of our target DNA of interest, what can we do with it? There are lots of ways to assay variation in an amplicon—one could digest it with restriction enzymes, for example, to look for RFLPs, which is much faster and easier than doing Southern blots, because after electrophoresis of the digested amplicon, the gel can simply be stained with a dye that binds to DNA and fluoresces under ultraviolet light (see Figure 7.8 for an example). Again, the fact that

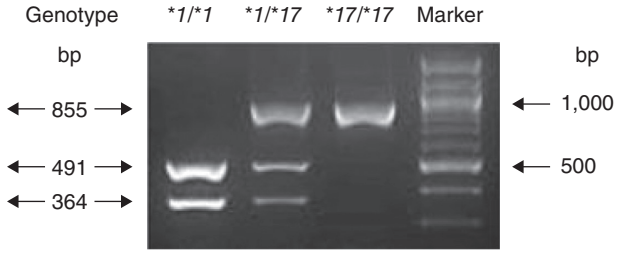


FIGURE 7.8

Example of the results of a PCR–RFLP experiment. In this example, there are two alleles, called *17 and *1; the *17 allele does not have a restriction site for a particular enzyme but the *1 allele does. When the polymerase chain reaction (PCR) product is digested with this restriction enzyme, individuals homozygous for the *17 allele have an intact 855 bp PCR product, individuals homozygous for the *1 allele have only 491 and 364 bp products (indicating the presence of the restriction site), and heterozygotes have all three bands. "Marker" indicates a molecular size marker, consisting of DNA fragments of known size. Reprinted with permission from Fukami, T., et al., "A novel polymorphism of human CYP2A6 gene CYP2A6*17 has an amino acid substitution (V365M) that decreases enzymatic activity in vitro and in vivo," *Clinical Pharmacology & Therapeutics* 76:519, 2004.

after PCR there are billions of copies of the target DNA compared to non–target DNA in the sample means that the dominant signal will come from the amplicon; everything else is just background noise. The DNA sequence of amplicons can also be determined quite easily, using methods described in the "DNA sequencing: the Sanger method" section. There are also a wide variety of methods for determining which nucleotide is present at a particular position, such as single-base extension assays, or for determining whether there are insertions or deletions of DNA in the target region, or for screening an entire amplicon for any new mutation, such as heteroduplex assays—too many methods to go into detail here! The main take-home message: PCR results in, for all practical purposes, a sample consisting only of the target DNA of interest, and so much of it that it can easily be manipulated further to look for variation in the target DNA.

The polymerase chain reaction was the brainchild of Kary Mullis, a chemist who at the time was working at the Cetus biotechnology company. According to his autobiography (entitled *Dancing Naked in the Mind Field*), Mullis came up with the idea one night in April 1983 while driving from San Francisco to his cabin in northern California. While it was a brilliant idea, it took a lot of effort from many other people at Cetus to get it to actually work. Nevertheless, in 1993 Mullis alone was awarded a Nobel Prize for inventing PCR—which brings up an ongoing issue with such prizes for which there is no clear answer, namely, who deserves more credit for an invention, the person who comes up with the idea, or the person who actually gets it to work?

Anyway, on one issue everyone agrees, namely, that PCR revolutionized many different fields of molecular genetics, including clinical diagnosis of DNA variants associated with disease; forensic DNA applications; and—of particular interest for this book—molecular anthropology. When I was carrying out my PhD research in the early 1980s on human mitochondrial DNA (mtDNA) variation (mtDNA will be explained in more detail in Chapter 9), we couldn't get enough mtDNA from a typical blood sample for our analyses—you'd have to bleed a person dry to get enough mtDNA, and it's hard to get volunteers for such studies! So we were forced to resort to placentas, and you can well imagine the difficulties involved in getting freshly frozen placentas from different populations. Plus, it took about 2 weeks to isolate the mtDNA from a placenta and several more weeks to carry out the analyses. But now, thanks to PCR, we can use not only blood samples but also plucked hairs, cheek swabs, and (our current favorite DNA source) saliva—for the latter, people simply spit into a tube with some chemicals to preserve the DNA, so sampling is fast, easy, and completely noninvasive, and the samples are

easy to transport as they don't need refrigeration. DNA purification from such samples takes a few hours, not weeks, because thanks to PCR, we don't need to purify mtDNA from all of the other DNA in order to analyze it, we simply purify all of the DNA at once. And the PCR itself is completely automated: add your DNA to a tube with the appropriate chemicals, primers, DNA polymerase, and nucleotides; put it in a thermal cycling machine that carries out the heating and cooling; push a few buttons and let it go—you barely have enough time for lunch before it's finished and you can continue with whatever further analyses you want to do. As we shall see in subsequent chapters, much of molecular anthropology is a numbers game; we want enough samples from each population to adequately characterize genetic variation in that population, we want enough populations to adequately characterize the variation among them, and we want to analyze enough genetic variation so that we can make accurate inferences about the history of these populations. So for molecular anthropology, PCR greatly facilitates getting the numbers. Moreover, as we shall see in Chapter 15, PCR basically enabled the whole field of ancient DNA.

DNA SEQUENCING: THE SANGER METHOD

A shortcoming of RFLP analysis is that it can detect only a subset of the mutations that occur, that is, mutations that create or destroy a restriction enzyme recognition sequence. Obviously, detecting some of the mutations is better than not being able to detect any at all, but the development of a rapid and efficient method for determining the entire sequence of a stretch of DNA, and hence all mutations that have occurred in that DNA, has (in combination with PCR) rendered RFLP analysis largely obsolete. This method is known as Sanger sequencing, named for the person who developed it in the late 1970s, Fred Sanger (Sanger and Coulson 1975; Sanger et al. 1977); he is the only person to win two Nobel Prizes in chemistry, one for DNA sequencing, the other for being the first to determine the amino acid sequence of a protein, insulin. The basic steps in Sanger sequencing are outlined in Figure 7.9 and rely on DNA synthesis initiated from a specific primer in the presence of modified nucleotides called **dideoxynucleotides**. These are analogues of the four usual nucleotide bases (A, T, G, and C) that can be incorporated into a growing DNA strand, but they prevent the DNA strand from further elongation—that is, they block further DNA synthesis—and hence the new DNA strand terminates at that position. So, if you include both normal nucleotides and dideoxynucleotides in the sequencing

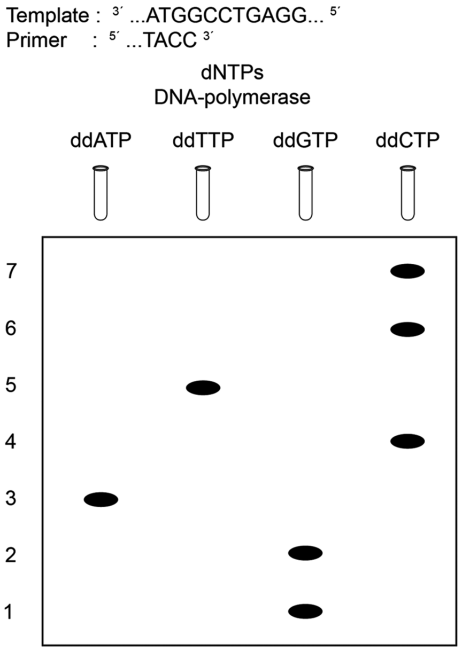


FIGURE 7.9

Principle behind Sanger sequencing with radioactively labeled bases. Shown at the top is part of a single-stranded template DNA sequence to which a radioactively labeled primer has annealed (the primer is usually about 30 bases long, so the complementarity between the template and the primer would extend further to the left than is shown). Four separate reactions are prepared, each containing template, primer, all four bases (dNTPs), DNA polymerase, and one dideoxy base (ddATP, ddTTP, ddGTP, or ddCTP). After the sequencing reactions are complete, the product from each tube is loaded into a separate lane of a polyacrylamide gel, subject to electrophoresis, and then dried and exposed to X-ray film. The idea is as follows: if you look at the first position at the 3′ end of the primer, which is where DNA synthesis will occur, the template is a C so a G will be added to the primer. If this is a dGTP, then DNA synthesis will continue to the next position. But if a ddGTP is added to the primer, DNA synthesis will stop because dideoxy bases are modified such that they can be added to a growing DNA strand, but no bases can be subsequently added to them (which is why another name for this procedure is chain termination sequencing). So all DNA molecules consisting of the primer plus one more base will appear in the ddGTP lane (position 1 on the left). Importantly, the ddGTP reaction tube also contains dGTP, so some DNA strands will incorporate dGTP and continue elongating. Since the second position in the template is also a C, DNA strands consisting of the primer plus two bases will also appear only in the ddGTP lane. Thus, by going from bottom to top, we can read off the DNA sequence as GGACTCC, which is exactly complementary to the template DNA sequence.

reaction, sometimes a normal nucleotide will be incorporated and the DNA strand will continue growing, but sometimes a dideoxynucleotide will be incorporated and the strand will terminate. The key point is that all of the strands that terminate at a specific position will have a specific dideoxynucleotide incorporated at that position, so if we have some way of knowing which dideoxynucleotide was incorporated, we then know the DNA sequence at that position. As originally developed, radioactively labeled nucleotides were used to visualize the DNA strands, so four separate reactions would be carried out, with each reaction containing all four nucleotides plus one dideoxynucleotide (Figure 7.9). These four reactions would then be electrophoresed side-by-side through thin gels made of polyacrylamide under appropriate conditions such that DNA strands that differ in length by as little as one nucleotide are separated. After electrophoresis, the gel is dried under a vacuum and then exposed to X-ray film; if everything has worked properly, then the DNA sequence can easily be read from the X-ray film (Figure 7.9).

Sanger sequencing was developed in the pre-PCR era and was originally carried out using DNA templates cloned into specialized **vectors** (either plasmids or viruses that infect bacteria) that had been modified to make them especially suitable for Sanger sequencing. Most sequencing that was carried out was so-called **shotgun sequencing**: take a sample of DNA, chop it up into lots of overlapping fragments (akin to blasting the DNA with a shotgun, hence the name), clone and sequence the fragments, and then use sophisticated computer programs to find overlaps among the DNA sequences, which are then used to assemble the "complete" DNA sequence. Pretty laborious, especially if you want to sequence an entire genome, as Sanger sequencing in those days typically gave DNA sequences around 400 bases in length. To sequence the entire human genome in this way would require more than 8 million clones, just to cover each position on average one time (and at that average coverage, you'd still expect about 1/3 of the bases to be missed!). So to guarantee that the majority of the positions in a DNA sequence are covered by shotgun sequencing, enough fragments have to be sequenced so that the average coverage is at least threefold. Despite these limitations, some smaller genome sequences, such as that of the bacterial virus phiX174 (about 5300 bases) and human mtDNA (about 16,500 bases), were determined in this way.

You might think that it would make more sense to come up with a strategy that would require sequencing each base just one time and then be done with it. But figuring out how to do so while knowing exactly where you are in the sequence was a daunting challenge, and in the meantime several technical advances allowed automation of Sanger sequencing, thereby making it all the easier to carry out shotgun sequencing. These included the introduction of nucleotides labeled with fluorescent dyes that could be detected with a laser (so no need to deal with nasty radioactivity) and the use of capillaries (thin tubes filled with a polymer) to carry out the electrophoresis, which were much faster and easier to work with than polyacrylamide gels. These, in turn, permitted the development of automated sequencing machines to carry out the capillary electrophoresis and laser detection. The output of such machines is a **chromatogram** (Figure 7.10), which can also be automatically exported to computers for further analyses. So, in the end, the human genome sequence was determined (largely) by the shotgun sequencing approach, using hundreds of such sequencing machines running night and day, along with sophisticated computational approaches to assemble all of the bits and pieces of shotgunned fragments into chromosome-length DNA sequences. And, of course, PCR was another major advance as it greatly simplified the preparation of specific targets for DNA sequencing. Methods were quickly worked out for reliably sequencing PCR amplicons directly—no more need to clone DNA into sequencing vectors. And for molecular anthropology, PCR-based sequencing was a dream come true, as it made it possible to investigate variation in the same DNA segment in different individuals and different populations.

NEXT-GENERATION SEQUENCING

Sanger-based sequencing, using automated machines based on capillary electrophoresis and fluorescent dye-labeled nucleotides, became a mainstay of DNA sequencing for well over a decade. The human genome sequence and other genome sequences were determined with this methodology, for example, and it is still widely used and considered the "gold standard" for DNA sequencing. However, beginning around 2005, new methods were introduced that utilize massively parallel processes to simultaneously sequence up to millions of DNA template molecules. These methods are referred to as **next-generation** or **high-throughput sequencing**, and they can generate enormous amounts of data at a fraction of the cost of Sanger-sequencing. For those of us who "grew up" with Sanger sequencing, the increased data output of the next-generation platforms is truly daunting—I recently came across a lecture on next-generation sequencing by Rob DeSalle from the American Museum of Natural History with the title "My Student Can Do My Ph.D. Thesis Research in 10 Minutes," which aptly describes how some of us feel! As I write this, there are several different next-generation

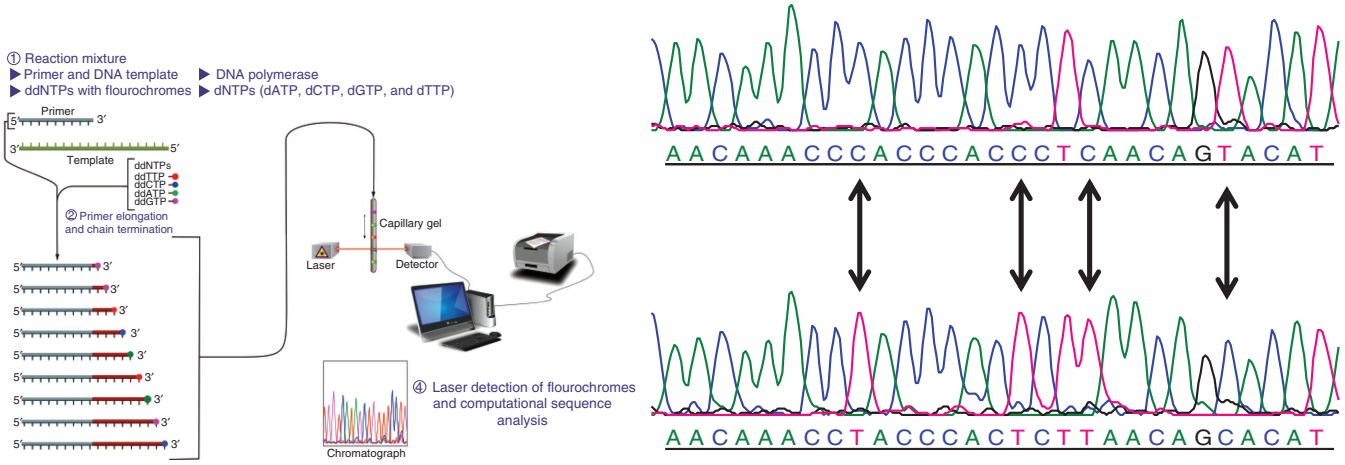


FIGURE 7.10

Left, the principle behind Sanger sequencing with dye-labeled terminating bases (ddTTP, ddCTP, ddATP, ddGTP). A reaction mixture is prepared containing the DNA template to be sequenced (e.g., a PCR product), a specific primer to initiate DNA synthesis, an enzyme to replicate DNA (DNA polymerase), and normal bases (dNTPs, where the N can be any of the four bases). DNA synthesis is initiated starting from the primer; if a dNTP is incorporated, the DNA strand keeps growing, but if a ddNTP is incorporated, the DNA strand stops at that position. The key point is that all DNA strands of a certain length will all have the same base at one end and hence will all have the same dye. After the reaction is done, the mixture is loaded onto a capillary gel, which separates the DNA strands by size, and a laser excites the dye and the signal is then recorded. The output is a chromatogram, where the peaks from left to right indicate the positions as the DNA strand gets longer, and the color of the peak indicates the base at that position. Right, an example of two chromatograms of mtDNA sequences from two individuals, with differences in the sequences noted. Left, reprinted with permission from, https://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg; right, courtesy of B. Pakendorf, reprinted with permission.

sequencing platforms available, with more on the horizon. The methods are changing so rapidly that it does not make sense to go into the details here; some details concerning current methods will be provided later on in the context of specific examples (e.g., Figure 9.3), and in any event easy-to-follow descriptions can be readily found on the Internet. The ultimate goal is to have a method that will generate a human genome sequence in a day or so for $1000 (by contrast, the initial human genome sequence took about 10 years and cost more than a billion dollars), and it seems likely that this goal will be met (or exceeded) by the time this book appears.

Some general comments here about the properties of current next-generation sequencing methods are, however, in order (and for a specific example, see Figure 9.3 later on in this book). These methods generally start by shearing a DNA sample to a specific average size (currently around 200–800 base-pairs, depending on the application), which is easily done by sonication—exposing DNA in a solution to ultrasound waves will fragment the DNA, and the longer the exposure, the shorter the DNA fragments. Specific DNA **adapters** are then attached to the ends of the fragments, which can then be used to PCR-amplify all of the fragments (the collection of DNA fragments with adapters attached is known as a **library**). So, a sequencing library typically will consist of millions and millions of DNA fragments, all with adapters at the ends, and is essentially an inexhaustible source of the DNA in the library (as the adapters can be used to PCR-amplify the library to make more of it). Through a variety of clever tricks, these fragments are physically separated from one another and sequenced in parallel, with several million fragments sequenced simultaneously. Processing all of the data from a single run of a next-generation sequencing machine can take several days and generate terabytes of data, so the computational demands of next-generation sequencing are not trivial.

The error rate associated with next-generation sequencing platforms tends to be quite high, as much as 10 times (or more) the error rate per base for Sanger sequencing. However, in general this is not a cause for concern, because in next-generation sequencing typically each base is sequenced—in shotgun fashion—on average tens to hundreds of times (or even more, depending on the application). The average number of times each nucleotide is sequenced is called the **coverage**, so a sequence with 30X coverage means that each nucleotide has been sequenced independently an average of 30 times. This repetition ensures that the

correct sequence will appear far more often than not, guaranteeing high accuracy—even higher than that for Sanger sequencing! What tends to be more of a problem for next-generation sequencing is that with the current platforms the read lengths are short, from a few tens to a few hundreds of bases. This, in turn, can make it quite difficult to assemble sequences accurately, especially in regions of repetitive DNA (i.e., repeated copies of the same DNA sequence—more on this later in this chapter). Some regions of the human genome simply cannot be accurately sequenced with current next-generation methods; for example, a "complete" human genome sequence determined to an average coverage of 30X should in theory not have any missing bases if the coverage truly was randomly distributed across the genome, yet in fact such sequences are actually "only" about 85% complete. Still, having the ability to rapidly and at low cost generate 85% complete human genome sequences—and, as we shall see later, generate high-quality sequences of specific target regions—means that next-generation sequencing is rapidly becoming the method of choice for most DNA sequencing applications, including many of particular interest for molecular anthropology.

TARGETING SINGLE DNA BASES: SNPs

Sometimes we are interested only in the genotypes of our samples at particular positions in a DNA sequence, rather than at all positions in the sequence. Often, this is because we have prior knowledge that these positions are variable, and so for a particular application it might be easier and/or cheaper to just focus on those positions. Single positions in a DNA sequence that exhibit variation in terms of the nucleotides present either in different individuals or in the two chromosomes within a single individual are known as **single nucleotide polymorphisms** or **SNPs** (pronounced "snips" by the cognoscenti) for short. There are a large variety of methods to determine genotypes of a sample of individuals for a particular SNP of interest. For example, if the SNP happens to either create or destroy a restriction enzyme recognition sequence, one could design a **PCR–RFLP** assay (cf. Figure 7.8): amplify a PCR product that contains the SNP of interest and digest the amplicons with the appropriate restriction enzyme. There are also Sanger-sequencing–based methods for sequencing just a single base position—these are so-called single base extension or "minisequencing" assays. It is also possible to combine or "multiplex" assays based on minisequencing (or other methods) for several different SNPs, thereby genotyping several SNPs in one individual in one assay. For example, useful multiplex assays have been developed for the major mtDNA and Y chromosome haplogroups, discussed in Chapter 9.

The real power of SNPs for molecular anthropology, however, arises from a technology that has been around for a decade or so that allows a very large number of SNPs to be genotyped simultaneously in an individual. There are a variety of ways to do this, but the general idea is that a DNA sample is labeled so that it can be detected later (e.g., with biotin) and hybridized to a so-called **SNP chip**, which is a microarray that contains synthetic DNA probes to the alleles for each SNP (Figure 7.11). Under the right conditions, DNA will hybridize only to the probes for the allele(s) in that sample—for example, if a SNP consists of either an A or a T at a particular nucleotide position, then DNA from AA homozygotes will hybridize only to probes to the A allele, DNA from TT homozygotes will hybridize only to the T allele, and DNA from AT heterozygotes will hybridize to the probes to both alleles (Figure 7.11). Knowing the exact position of the probes for this SNP on the microarray then permits the genotype of a DNA sample to be determined for this SNP by simply determining which probes give off a signal after hybridization and staining for bound DNA (e.g., DNA that contains biotin). And by placing probes to lots and lots of SNPs on the microarray, the genotypes for all of these SNPs can be determined with just one simple assay.

SNP chips were originally designed in the early 2000s for a few tens of thousands of SNPs but nowadays are available for millions of SNPs. They were designed for disease-association studies; the idea is that if there exists a particular genotype at a particular locus that enhances risk for a particular disease, then by genotyping lots of SNPs in a group of patients with the disease and the same SNPs in controls without the disease, any SNP allele that is associated with higher risk for the disease will be at higher frequency in the patients than in the controls. Note that "associated with higher risk for the disease" does not mean that the SNP necessarily plays any role in actually causing the disease, as any SNP that is in **linkage disequilibrium** (explained in Chapter 9) with a mutation that actually does cause the disease will show an increased frequency in the patients relative to the controls. Thus, a key feature of SNP chips is to have enough SNPs that are spaced closely enough around the genome to be able to "tag" any disease-causing mutation via linkage disequilibrium (this is the idea behind so-called **tag SNPs**). Note also that to determine when a particular SNP allele is at significantly higher frequency in patients than in controls, one uses the distribution of allele frequency differences between patients and controls at all of the SNPs on the chip—thus, the SNP chips automatically provide all of the necessary background information to identify disease-associated SNPs.

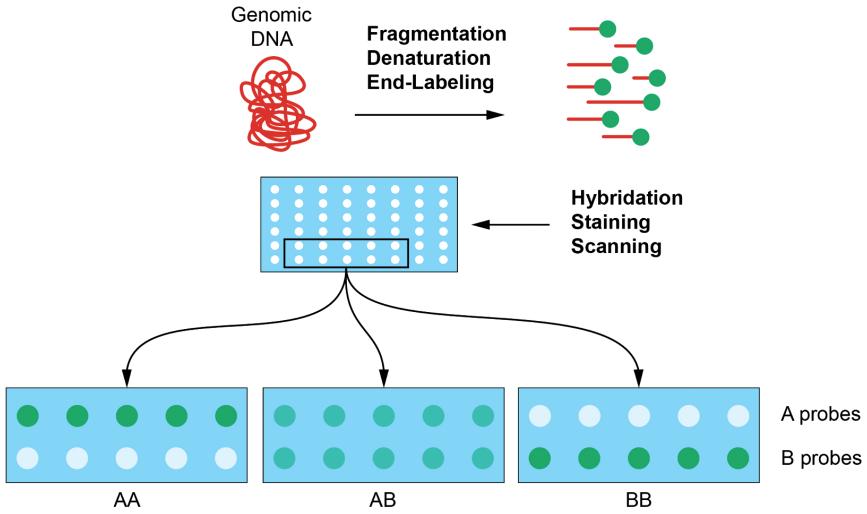


FIGURE 7.11

Principle behind genotyping SNPs by hybridization to DNA probes on an array (so-called "SNP chips"). Purified genomic DNA is fragmented, denatured to single strands, labeled with a fluorescent dye, and hybridized to an array containing several probes to each of the two alleles at a SNP (alleles A and B). The hybridization is done under very stringent conditions, so that DNA will hybridize only to probes that match the DNA sequence perfectly. After staining for fluorescence, washing, and scanning, the expectation is that homozygotes for the A allele will show fluorescence only for the A probes, homozygotes for the B allele will show fluorescence only for the B probes, and heterozygotes will show fluorescence for both sets of probes. The use of several probes for each allele at a SNP enables more accurate SNP genotype calls.

The convenience of SNP chips led to an entire industry of **genome-wide association studies** (or GWAS, pronounced "gee-was"). Such studies produced some spectacular successes in identifying key genes involved in susceptibility to various diseases, such as the first GWAS in 2005 (Klein et al. 2005), which identified two mutations that increased susceptibility to age-related macular degeneration in a gene that nobody suspected had anything to do with the disease, or a later study in 2007 (with more than 14,000 patients!) that identified several novel genes involved in susceptibility to heart disease, diabetes, and rheumatoid arthritis (The Wellcome Trust Case Control Consortium 2007). But there also have been numerous failures of the GWAS approach to come up with significant candidates, and even the disease-susceptibility candidates that have been identified typically increase the risk of disease by just a few percent. Something else must be going on—what that might be is currently a source of considerable debate, but it seems likely that many complex diseases (such as diabetes or heart disease or mental disorders) are influenced by mutations at many different genes. Thus, you and I may suffer from the same disease, but the underlying cause is different mutations in different genes. And, of course, with complex diseases it is not a simple matter of "if you have the allele, you get the disease"; the environment also plays a crucial role in an individual's disease risk, as was discussed back in Chapter 1. Many investigators are turning to sequencing complete genomes from patients in the hopes of thereby identifying the underlying susceptibility mutations. However, the success of this approach is far from assured, as any given individual might carry upward of a million genetic variants, so figuring out which ones (if any) might be involved with a particular disease is a daunting task.

Anyway, the GWAS approach may seem rather remote from molecular anthropology, but it has been used to identify signatures of selection in human populations (discussed in Chapters 17 and 18), as well as to identify genes underlying phenotypic traits of anthropological interest (such as skin pigmentation, discussed in Chapter 20). Furthermore, the SNP chips originally developed for GWAS are being increasingly used to study human population history, as they provide a relatively fast, easy, and inexpensive (well, a few hundred dollars per sample) means of obtaining genotypes for hundreds of thousands to millions of SNPs. As we shall see in Chapter 11, the resulting data are allowing all sorts of new analyses that previously were not feasible and providing all sorts of new insights into human population history.

But alas, there ain't no such thing as a free lunch (or TANSTAAFL, as one of my professors used to say), and using SNP chips for population history does not come without drawbacks. The chief drawback has to do with how the SNPs on the chips were chosen—they have to be polymorphic, of course, and they should be

reasonably well-spaced around the genome. In selecting SNPs for their chips, the companies involved naturally relied on databases that were available, and most of these were heavily biased toward SNPs discovered in individuals of European ancestry. Thus, SNP chips overestimate the amount of genetic variation in general (because they focus only on variable positions), and in particular they overestimate the amount of genetic variation in populations of European ancestry relative to non-European populations. This is an example of **ascertainment bias**, and a recurring theme that will come up several times in this book is that ascertainment bias can seriously compromise the results of some analyses unless one takes care to correct for it. Fortunately, methods do exist that either correct for, or are not so terribly influenced by, ascertainment bias, so SNP chips currently continue to be an enormously fruitful source of new insights into human population history. Still, as sequencing costs continue to go down, most of us expect that SNP chips will be eclipsed by partial or complete genome sequencing (which does not suffer from ascertainment bias, as sequencing reveals all of the variation present in the genomic region sequenced), and probably sooner rather than later.

VARIATION IN LENGTH

So far we have been concerned with variation that consists of substitution of one nucleotide for another at a specific position in a DNA molecule, for example, SNPs. However, there can also be variation in the length of a specific segment of DNA, and there are many different kinds of length variation that can occur. The simplest such variation is insertion or deletion of one or a few nucleotide(s) at a single position, and such variants are imaginatively called **indels** (for **in**sertion–**del**etion). Indels are generally detected by the usual methods for detecting nucleotide sequence variation, including RFLPs in the old days and DNA sequencing nowadays. The importance of indels in human evolution is an area of active research, as while they tend to be somewhat less numerous than nucleotide substitutions (e.g., the chimpanzee and human genome sequences differ by about 35 million nucleotide substitutions and 5 million indels), they could have significant functional consequences.

Other length variation in genomes usually involves repeated copies of the same DNA sequence, or **repetitive DNA**, and we can distinguish between DNA repeats scattered across the genome (**interspersed repeats**) versus variation in the number of copies of a DNA repeat at a specific locus (**tandem repeats**). We will first go over interspersed repeats and then discuss tandem repeats.

Interspersed Repeats

You've already seen one example of interspersed repeats, and that is the Alu family discussed in Chapter 6. Recall that Alu elements are short (~300 bp) elements, derived from a so-called "master" element that periodically makes a copy of itself that inserts into a new location in the genome. This process has been ongoing since the origin of primates, with the result that most humans will have the same Alu elements at the same chromosomal locations—such Alu elements have become fixed in the human species. However, some Alu elements have inserted so recently that not all humans have an Alu element at the specific chromosomal location. Polymorphism for the presence or absence of an Alu element is known as an **Alu insertion polymorphism** (AIP), and there are three possible genotypes for an AIP: homozygous for the presence of the Alu element; homozygous for the absence of the Alu element; and heterozygous (one chromosome has the Alu element and one chromosome lacks the Alu element). These genotypes can easily be distinguished via a PCR assay with primers that flank the insertion point for an AIP, as amplicons with the Alu element will be about 300 bp longer than amplicons lacking the Alu element, and this size difference can easily be visualized by simple agarose gel electrophoresis of the amplicons (Figure 7.12). In addition to being very simple to assay, AIPs have other desirable properties: they represent unique events during human evolution, as the chance that an Alu element would insert independently in different individuals at the exact same position in the genome is negligible; they are stable markers, as deletion of Alu elements only rarely occurs, and when it does a footprint of the deletion event is evident since the Alu element is either only partially deleted, or some of the flanking human DNA sequence is also deleted; and the ancestral state is known to be the absence of the Alu element while the derived state – i.e., the direction of mutation – is the presence of the Alu element at a particular location in the genome (knowing which is the ancestral allele and which is the derived allele for a polymorphism is crucial for some analyses). Alu insertion polymorphisms and other polymorphisms involving the presence/absence of other interspersed repeat elements have been valuable and informative genetic markers for studies of human population history.

Tandem Repeats: Minisatellites

The other major type of length variation involves variation in the number of copies of a tandemly repeated segment of DNA. The first such variation to be discovered, in 1980 (Wyman and White 1980), involved a random piece of human DNA isolated from a collection of human DNA fragments that had

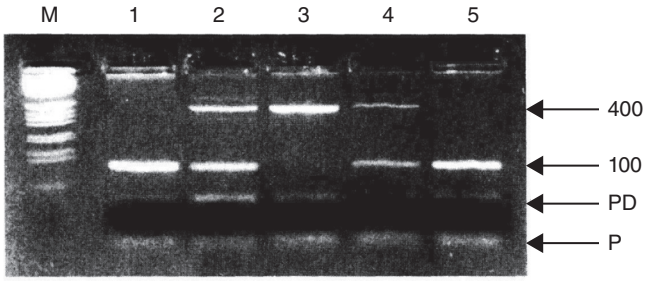


FIGURE 7.12

Visualization of the results of polymerase chain reaction (PCR) using primers that flank a polymorphic Alu insertion locus. The photograph shows an agarose gel after electrophoresis of the PCR products followed by staining the gel with ethidium bromide, which binds to the DNA and fluoresces under ultraviolet light. M is the molecular weight size marker, consisting of DNA fragments of known size. Lanes 1 and 5 are from individuals homozygous for the absence of the Alu element and hence show a ~100 bp product. Lane 3 is from an individual homozygous for the presence of the Alu element, which is about 300 bp, and so this individual shows a ~400 bp product. Lanes 2 and 4 are from individuals heterozygous for the presence of the Alu element at this locus and hence show both the 100 bp and 400 bp products. Thus, the genotypes can be easily determined simply by agarose gel electrophoresis of the PCR products. PD is a so-called primer dimer band, reflecting primers that sometimes anneal to one another instead of to the template DNA and thus produce a spurious product that is smaller than the expected products and hence does not interfere with genotype determination; P is a band reflecting excess primer DNA. Reprinted with permission from Perna, N., et al., "Alu insertion polymorphism: a new type of marker for human population studies," *Human Biology* 64:641, 1992.

been cloned into a bacteriophage to make a human genomic library. When this random piece of DNA was used as a probe in Southern blots with DNA from different humans, it revealed a large number of alleles that were inherited in simple Mendelian codominant fashion. Further investigation showed that the alleles differed in the number of copies of a core sequence (Figure 7.13 shows an example), with longer alleles having more copies. This was the first variable number of tandem repeats locus described in humans.

When Sir Alec Jeffreys, a British geneticist, saw this, he thought that there should exist other such examples of highly variable, tandem repeat loci in humans, and he set out to find them. He quickly isolated other such probes but found to his astonishment that when one of them was hybridized to a Southern blot of human DNA, he obtained a large series of bands (Figure 7.14).

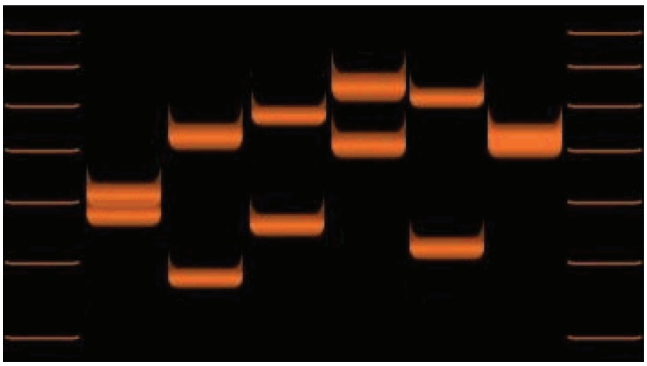


FIGURE 7.13

Results of gel electrophoresis of the polymerase chain reaction products for a VNTR locus. The flanking lanes are the molecular weight size markers, while the six lanes in between illustrate different genotypes. All individuals have two bands, indicating that all individuals are heterozygous for two alleles with different numbers of repeats, as is commonly observed for highly polymorphic VNTR loci. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:D1S80Demo.png).

The reason for these multiple bands is that the particular core sequence included in the probe is present at several hundred locations throughout the genome, with a variable number of tandem repeats of the core sequence at each location. Such core sequences are known as **minisatellites**—the terminology is based on satellite DNA, which is highly repetitive DNA that got its name because it forms a discrete secondary or "satellite" zone, separate from the zone where most genomic DNA occurs, when DNA is centrifuged in a density gradient at high speed (which used to be a common method of purifying DNA). Minisatellites have smaller repeat structures than satellites, hence the name.

Most importantly, the banding pattern revealed by the Southern blots differed for each individual, even among close relatives (though close relatives clearly shared more bands in common than unrelated individuals). Sir Jeffreys had thus discovered **DNA fingerprints**, that is, the use of DNA to uniquely identify an individual (Jeffreys et al. 1985). And not long after they were first described, DNA fingerprints were used by Sir Jeffreys in the first application of DNA typing to a forensic case—you can read all about it in a book by noted crime novelist Joseph Wambaugh, called *The Blooding: The True Story of the Narborough Village Murders*. Two teenage girls had been raped and strangled, in 1983 and 1986, and a prime suspect had confessed to the second killing but denied responsibility for the first. When DNA fingerprinting was applied to semen stains from both rapes and to a blood sample from the

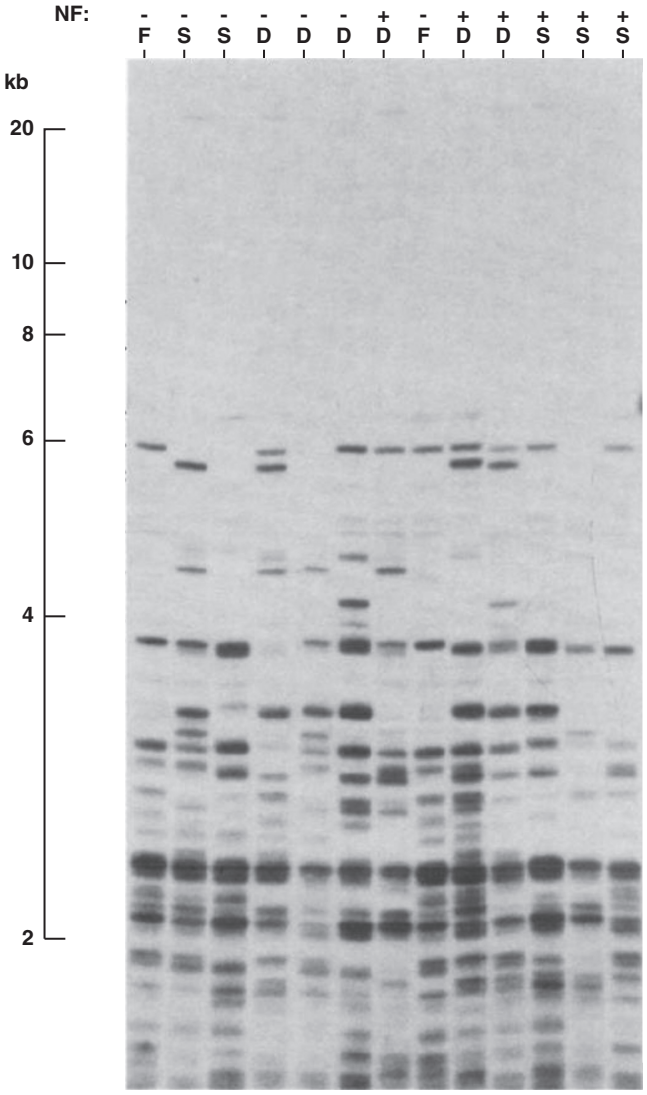


FIGURE 7.14

Autoradiogram of DNA fingerprints obtained from several individuals. Each column (lane) is a DNA extract from a single individual that was subject to gel electrophoresis and then Southern blotting (as outlined in Figure 7.5) with a DNA probe that detects many related DNA sequences across the genome that vary in length. Even though many of the individuals in this autoradiogram are related and share bands, each individual has a unique DNA fingerprint. Reprinted with permission from Jeffreys, A., "Highly variable minisatellites and DNA fingerprints," *Biochemical Society Transactions* 15:309, 1987.

suspect, the results indicated that the culprit was the same in both rape–murders—but the culprit was not the prime suspect, even though he had confessed to one killing! The police then resorted to the extraordinary measure of requiring all "eligible" males in the vicinity of the murders to provide a blood sample for DNA fingerprinting, in an effort to identify the culprit. If you are from the United States, you might be wondering how on earth the police could get away with such a demand, as surely there would be public outrage (and a flood of lawsuits), but in this respect, the United States and Britain evidently differ, as nobody complained, and samples from about 5000 men were processed in about 6 months. You might also wonder why on earth the culprit would voluntarily provide a blood sample that would implicate him in the crimes, and here you would be correct, as none of the DNA fingerprints obtained in this unprecedented screening matched that of the culprit. And yet in a way the screening did result in the capture of the culprit, because later a man was overheard bragging in a pub that he had been paid by a friend (a local baker) to pose as him and provide a blood sample. The police quickly located the baker, obtained a blood sample, and the baker's DNA fingerprint matched that of the culprit. Based on this evidence, the baker was convicted of the rape–murders of the two girls. Thus, this first use of DNA in a forensic case nicely illustrates the power of DNA fingerprinting to both exonerate the innocent (e.g., the individual who had even confessed to one of the crimes) and incriminate the guilty. In the absence of the DNA evidence, it is quite likely that the first prime suspect would have been convicted of at least one of the murders, and the true murderer would never have been apprehended.

DNA fingerprinting has seen numerous applications, not just in forensic cases but also in paternity testing, immigration disputes (where a legal immigrant to the United Kingdom wants to bring in a family member, and there is a dispute as to whether the individual in question is really a family member or not), determining whether twins are dizygotic (arising from separate eggs) or monozygotic (arising from a single egg), monitoring tissue transplants, and even in wildlife and conservation biology (e.g., determining whether or not a particular specimen is from an endangered species). However, DNA fingerprinting hasn't had much of an impact on molecular anthropology, because the DNA fingerprints are so variable from individual to individual and also because of difficulties in determining whether what appears to be the same band in unrelated individuals is really the same band or not. Indeed, I spent a few months when I was a graduate student working with DNA fingerprints from individuals from different populations but ended up abandoning the effort because of the above difficulties (alas, not all good ideas work out!).

Tandem Repeats: Microsatellites

Another development that is closely related to DNA fingerprinting but has had a much bigger impact on

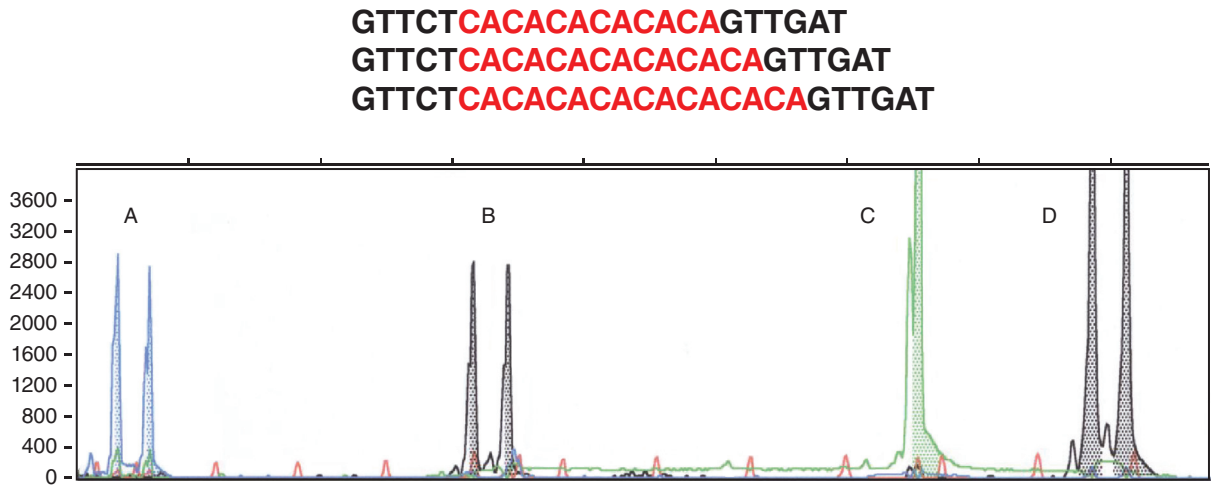


FIGURE 7.15

Top, sequence of an STR locus, showing variation in the number of CA repeats. Bottom, electropherogram for 4 STR loci (A–D) with polymerase chain reaction (PCR) products of different lengths. The polymerase chain reaction was carried out simultaneously for all four loci (i.e., with eight primers in the PCR) in a single DNA sample, with each primer pair labeled with a different fluorescent dye. The products were then separated by capillary electrophoresis and visualized using a laser to detect the fluorescence. The small red peaks are a molecular weight size marker (A has the smallest size fragments, D the largest size fragments), and the Y-axis is a measure of how much fluorescence was observed. The presence of two peaks for each locus indicates heterozygosity for each locus in this individual.

molecular anthropology is the PCR-based genotyping of **short tandem repeat** (STR) loci or **microsatellites** (based on the satellite–minisatellite nomenclature; microsatellites have even shorter repeat units). These are tandemly repeated copies of a 2–6 bp sequence and are quite frequent in the human genome (the most common are CA dinucleotide repeats, consisting of the sequence CA repeated a variable number of times), with the typical STR locus having several alleles that differ in the number of copies of the repeat sequence (Figure 7.15). It turns out that STR loci evolve by a very different mutational mechanism than nucleotide substitutions in DNA sequences. Typically, a new mutation at an STR locus will consist of the gain or loss of one repeat unit (e.g., an allele with 12 CA repeats will mutate to 11 or 13 CA repeats) due to slippage of the DNA polymerase during DNA replication—the DNA polymerase basically loses track of how many repeats there are. Mutation rates for STR loci are thus typically several orders of magnitude larger than nucleotide substitution rates, which accounts for why STR loci are so polymorphic. New mutations practically always involve the loss or addition of just one repeat unit; this means that STR loci evolve under a **stepwise mutation model** (i.e., one "step" or repeat unit at a time), which in turn necessitates somewhat different analytical methods. For example, stepwise mutations have a high probability of occurring independently in different individuals—if you and I each have an allele with seven repeats, then under the stepwise mutation model, if a new mutation arises in my child and also in your child, there is a 50% chance that our children will have the same mutations (both lost a repeat, or both gained a repeat). So, STR loci are a prime example of violating the infinite alleles model that we discussed back in Chapter 5. Furthermore, it turns out that the probability of an STR allele mutating is dependent to some extent on the number of repeats in that allele; alleles with more repeats are more likely to lose or gain a repeat than alleles with fewer repeats, probably because the greater the number of repeats, the more likely it is that the DNA polymerase will lose track of how many there are. So, not everybody has the same chance of a mutation when having offspring, which further complicates things.

Short tandem repeat loci came into their own as valuable genetic markers when the geneticist James Weber demonstrated in 1989 that they are highly polymorphic, codominant markers that can easily be genotyped via PCR using primers to the unique sequence that flanks the repeat region, thereby producing amplicons that differ in length according to the number of repeats (Weber and May 1989). While there are various methods for determining the length of the amplicons, the most widespread technology involves labeling amplicons with fluorescent dyes, followed by capillary electrophoresis and laser detection of the amplicons; by making use of different dyes and

amplicons of different lengths, several STR loci can be assayed in one go (Figure 7.15).

Weber (and others) quickly realized the value of such highly polymorphic markers for mapping disease genes—that is, determining the genomic region where a disease locus is found. Recall from Chapter 1 the discussion about linked loci, with the example of the Rh blood group and elliptocytosis loci. In this example, the Rh blood group can be thought of as a marker locus and the elliptocytosis locus as a disease locus; if we know where in the genome the Rh blood group locus is located, then the fact that the elliptocytosis locus is linked to it means that the elliptocytosis locus must be located near the Rh blood group locus, on the same chromosome. Recall also that in order to determine whether two loci are linked or not, one parent must be heterozygous for both loci. So the idea is that with a lot of highly variable marker loci, individuals will usually be heterozygous at most of them. And if we also know where in the genome these marker loci are located, then we can genotype them in families segregating for a disease of interest, see if any of the marker loci are linked to the disease locus, and if so, thereby identify a particular region on a particular chromosome where the disease locus is located. This, in turn, can facilitate the identification of the particular gene responsible for the disease, which can lead to new insights into what actually goes wrong in a particular disease, and may even lead to new treatments. Weber assembled a panel of several hundred STR loci with known chromosomal locations that effectively covered the entire genome and set up a genotyping service at the Marshfield Clinic in Wisconsin that enabled the genetic mapping of numerous disease-susceptibility loci.

Short tandem repeat loci also revolutionized the field of forensic genetics, as they are much easier to work with and analyze than DNA fingerprints. This is especially the case for samples with degraded and/or low amounts of DNA, which is frequently the case with crime scene samples, and DNA evidence has been successfully recovered from hairs, bones, cigarette butts, postage stamps—and, in a nice twist, even from fingerprints! A standardized set of 13 STR loci has been developed for forensic casework; the chance that two different individuals would have the same genotype at all 13 loci (or "profile") is essentially zero (except, of course, for identical twins—so if your genotype matches that of a crime scene sample, there is always the "evil twin" defense to fall back on!). In addition to identifying the perpetrators of crimes, DNA evidence—mostly based on STR genotyping of old crime scene samples—has been used to exonerate more than 250 people in the United States who were convicted of crimes they did not commit, including some who were facing the death penalty.

To further aid in the forensic use of DNA evidence, in the early 1990s, the FBI set up a database of STR profiles called the Combined DNA Index System (CODIS). The idea is to have a database of STR profiles from convicted felons, so that someone who later commits another crime and leaves a DNA sample at the crime scene can be quickly identified and apprehended. The CODIS database currently consists of profiles from more than 10 million individuals and has amply proved its value in identifying perpetrators—thousands and thousands of cases have been solved with the aid of CODIS, and CODIS has been featured on popular crime shows on television such as the CSI series. However, DNA databases are not without controversy over privacy concerns. In the United States, it is up to individual states to decide which crimes merit taking a DNA sample and putting the STR profiles into CODIS, and some states collect profiles not only from those convicted of a violent crime but also from those convicted of relatively minor crimes (such as passing a bad check) or even from arrested suspects who are later found innocent. Some go so far as advocating putting profiles from *everyone* into CODIS, reasoning that after all, if you never commit a crime, what do you have to worry about? Others see this as an extreme violation of the principle that people are presumed innocent until proven guilty and a gross intrusion of the government into personal privacy. Still, more and more countries are adopting DNA databases, and given how useful they are, such databases are undoubtedly with us to stay in some fashion or another—but what fashion that takes is up to the educated public (you, for example) to decide.

Anyway, STR loci have also had a profound impact on molecular anthropology due to the same features that made them attractive for linkage studies: namely, it is relatively easy to collect genotypes from a large number of markers that are highly polymorphic and hence are highly informative. Initial studies of STR variation were based on relatively small numbers of loci, but this changed when the Marshfield Clinic offered large-scale STR genotyping services to the human genetics community, not just for disease studies but also for population variation studies. Several important studies of genome-wide variation in various human populations were carried out that made use of this service (e.g., Friedlaender et al. 2008; Rosenberg et al. 2002; Tishkoff et al. 2009). In fact, it is only relatively recently, with the development of SNP-chip technology, that SNPs have overtaken STRs as the current methodology of choice for studies of genome-wide variation (and, keep in mind, next-generation genomic DNA sequencing will undoubtedly soon overtake SNP chips). Short tandem repeat loci have also proven especially useful in studies of human Y

chromosome variation, where initially there wasn't much DNA sequence variation to be found. These Y-STR loci can be used to estimate the ages of particular Y chromosome mutations of interest, and to investigate the paternal history of human populations, which can be particularly insightful when compared to the maternal history as revealed by mtDNA analyses; we'll see examples in Chapters 16 and 19.

One final point about STR loci: the widespread use of STR loci in forensics—especially the 13 loci that make up the core CODIS STR profile—has led to the development of commercially available kits that make it very quick and easy to carry out the genotyping for these loci. This, in turn, has tempted some investigators to use these loci to study and make inferences about population history. But an important caveat is that the commonly used forensic STR loci were selected specifically for forensic use because they show lower-than-average genetic differences between populations. Genetic differences between populations are of concern when figuring out the probability that a particular STR profile will be observed in another individual from the same population. To estimate this probability, you have to figure out which population is appropriate, and the concern is that if STR profiles differ a lot between populations, then using the incorrect reference population may give you a wrong probability value. Using loci that show small genetic differences between populations makes this less of a concern, because the probability values are similar regardless of the reference population used, and hence such STR loci are preferred for forensic casework. But this also means that if one uses such STR loci in molecular anthropological investigations, one will not get accurate estimates of genetic differentiation between populations—and, not surprisingly, such studies do tend to find smaller genetic differences among populations than revealed by other genetic markers. As emphasized in Chapter 9, the choice of which genetic markers to study should be dictated by the questions of interest, not by the availability of a kit that makes it easy to carry out the genotyping.

Tandem Repeats: Copy Number Variants

The final class of tandem repeats that deserves mention is **copy number variants**, or CNVs for short. These are arbitrarily defined as segments of DNA more than 1 kb in length that (as the name suggests) are present in different copy numbers in different individuals. They can be either missing entirely (deleted) or present in more than one copy (duplicated). The existence of CNVs was first detected with the completion of the human genome sequence in the early 2000s and initially relied upon cumbersome cytogenetic techniques such as fluorescent in situ hybridization, in which fluorescent-labeled DNA probes are hybridized to chromosomal preparations and the actual binding site(s) of the probes to a specific chromosomal region visualized under a microscope. Nowadays, CNVs can be detected from SNP chips, as the hybridization of a DNA sample to the probes on the chip is not only qualitative (to detect the SNP) but also quantitative in that the amount of fluorescent signal will be proportional to the amount of DNA that hybridized to the corresponding probe. Thus, samples with either more or fewer copies of a DNA segment will show correspondingly more or less signal for all probes in that segment (Figure 7.16). In fact, some commercially available SNP chips now contain probes specifically designed to enable detection of known CNVs. Next-generation sequencing can also readily detect CNVs, as the number of sequencing reads that map to a specific genomic region will depend on how many copies of that region are present in the sample sequenced. As with SNP-chip hybridization, a systematic decrease or increase in the number of reads mapping to a specific genomic region identifies a potential CNV.

Copy number variants are quite common in the human genome. They range from 1 kb (by definition) up to several million bp, and more than 20,000 CNVs have been identified that encompass around 20% of the human genome. Initially, it was hoped that CNVs might account for the genetic basis of complex diseases that could not be accounted for by SNPs, but further studies have alas so far failed to find any especially significant role for CNVs in complex diseases. However, the role of CNVs in human evolution and how they vary among human populations is just beginning to be systematically investigated (e.g., Sudamant et al. 2015), so stay tuned for further developments.

OTHER STRUCTURAL VARIATION

Copy number variants are one example of so-called structural variation (i.e., variation involving large chunks of DNA or chromosome segments). Other examples include **segmental duplications**, which basically are CNVs that involve duplications of a genomic region that have become fixed between species. It has been estimated that about 2.7% of the human and chimpanzee genomes differ by segmental duplications (Cheng et al. 2005), which is more than double the amount of single nucleotide differences between humans and chimpanzees (about 1.2%). Since these can involve several genes, and moreover can alter patterns of gene expression, they are interesting candidates for association with the phenotypic

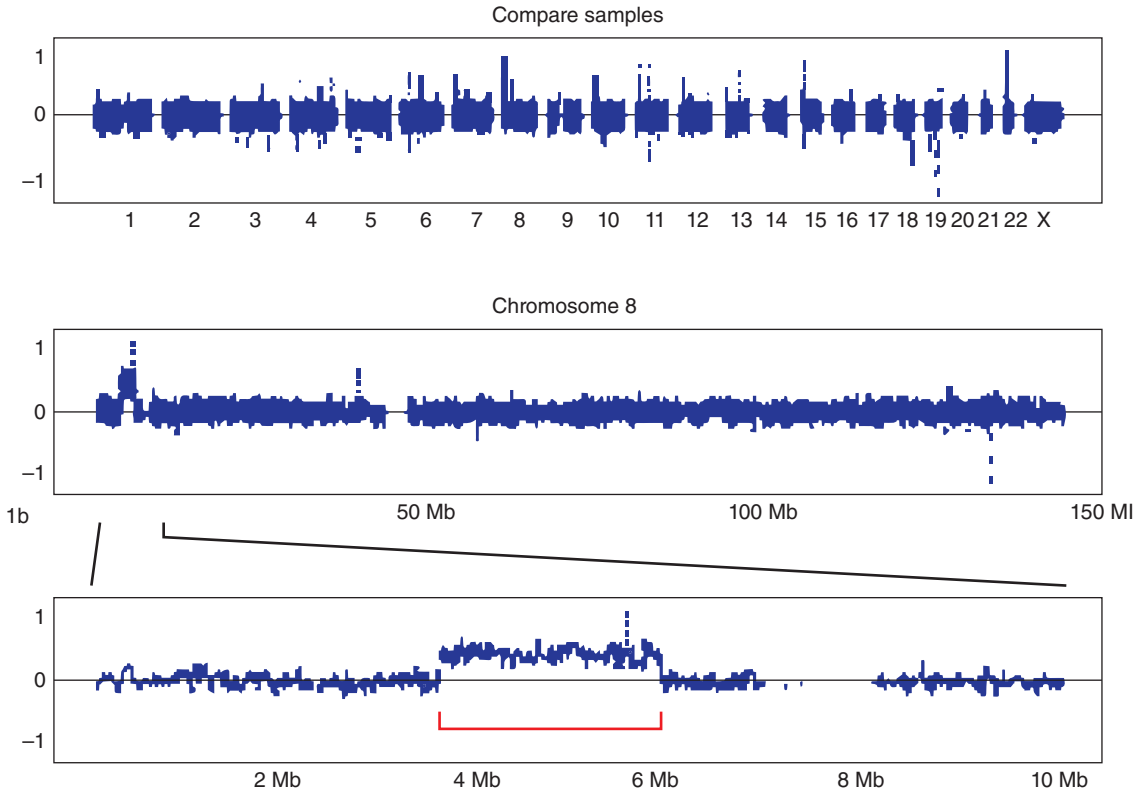


FIGURE 7.16

How SNP chips can be used to detect copy number variation (CNV). Top panel: Each blue dot is a result from one of more than 500,000 probes from across the genome (chromosomes indicated at the bottom of the panel), showing the (normalized) intensity of hybridization of one test sample compared to a reference sample. Normalized intensity values are expected to be around zero; large differences from zero indicated either more or less intensity in the test sample, which then reflects either more or fewer copies of that DNA sequence in the test sample. The middle panel shows an expanded view of chromosome 8, with a region near the beginning that shows significantly higher intensity in the test sample, while the bottom panel shows an expanded view of this region and indicates that a ~2 Mb (million base pair) region of chromosome 8 is duplicated in the test sample. Reprinted with permission from Redon, R., et al., "Global variation in copy number in the human genome," *Nature* 444:444, 2006.

differences between us and chimpanzees—altering the copy number of several genes at once could have important functional consequences. **Inversions** are another type of structural change that do not involve deletions or duplications but rather simply reversing the order of genomic regions. For example, if we have genomic segments in the order A-B-C-D-E, an inversion involving segments C and D would have the order A-B-D-C-E. Inversions are not easy to detect (unless they are big enough to visualize cytogenetically) as they don't alter the amount of DNA but just the relative order of DNA sequences along a chromosome, but they can be identified by careful analysis of full genome sequences for the inversion breakpoints. The human and chimpanzee genome sequences differ by nine large inversions and perhaps as many as 1500 smaller inversions that cover about 5% of the genome (Feuk et al. 2005), and a few polymorphic inversions have been identified in humans. Again, the functional significance (if any) of these structural changes remains to be seen, but one potentially important aspect of inversions is that they tend to suppress recombination, as recombination within an inverted segment in an individual heterozygous for an inversion leads to DNA duplications and deletions. In other organisms such as fruit flies, polymorphic inversions have been associated with combinations of interacting genetic variants that are selectively advantageous when kept together (so-called "co-adapted gene complexes"); whether or not this also holds for humans is an intriguing possibility that remains to be seen.

CONCLUDING REMARKS

As this chapter has shown, there is a huge variety of genetic markers that have been (or could be) used in

molecular anthropology studies, with various properties that impact how suitable they are for addressing particular questions of interest. A related concern is what region(s) of the genome to sample, as this will also impact the choice of genetic markers, that is, the methods that one uses to survey genetic variation in the genomic region(s) of interest. So, we need to discuss the different properties of different regions of the genome, with an eye as to how these properties might influence our results. But first, a word or two about sampling individuals and populations for molecular anthropology studies.

CHAPTER 8

SAMPLING POPULATIONS AND INDIVIDUALS

Some of the most important, yet least-appreciated, aspects of molecular anthropological studies are issues related to sampling. How does one choose which populations to study and which individuals to sample? And how does one choose which genes or DNA regions to study? Unfortunately, it all too often seems the case that little or no thought has gone into these questions. And yet, the choice of populations and/or DNA regions to sample can profoundly influence the outcome of a study. In this chapter, therefore, we will consider issues related to the sampling of individuals from populations, while in the next chapter we discuss sampling of DNA regions.

SAMPLING POPULATIONS: GENERAL ISSUES

Consider, by way of analogy, the rainbow in Figure 8.1. Suppose we wish to analyze the various colors that make up the rainbow, and we start by sampling the three parts indicated in the figure. We would then conclude that a rainbow consists of three very different colors—red, blue, and yellow, the primary colors—with easily distinguished properties. But suppose we instead take many samples across a segment of the rainbow (also as indicated in Figure 8.1); we would then conclude that a rainbow consists of a gradient of colors, one blending into another. While on the one hand these would seem to be contrary views, on the other hand, in some sense both views are correct. Part of a rainbow does indeed consist of discrete colors that can be readily distinguished from one another, but the rainbow as a whole shows continuous variation across the visible spectrum of light.

Sampling of human populations for genetic variation studies mirrors this analogy. Some studies have focused on just a few populations from locations that are quite distinct. For example, the first phase of the international HapMap project (described in more detail in the next chapter) included samples from just three populations: Europeans (actually, European–Americans from Utah); East Asians (Han Chinese from Beijing and Japanese from Tokyo, usually grouped together into one population); and Africans (Yoruba from Nigeria). Note that the part of Africa that is south of the Sahara is sometimes referred to as sub-Saharan Africa, as distinguished from northern Africa—in this book, Africa should be understood to refer to sub-Saharan Africa unless northern Africa is explicitly stated. Genetically, these three populations can be easily distinguished, and some studies based on HapMap or similar sampling schemes have concluded, either explicitly or implicitly, that the human species, therefore, consists of discrete groups that can be readily distinguished genetically. But this is no more correct than concluding that a rainbow consists of discrete colors, based on sampling the three primary colors. Indeed, as will be discussed in more detail in Chapter 14, in general, genetic variation is continuous or **clinal** across the range of human populations, rather than organized into discrete groups with defined boundaries between them. This is not to deny the existence of genetic differences between human populations (although, as we shall see, such differences are far outweighed by the genetic similarities among human populations), and it is certainly just as correct to say that we can distinguish the three HapMap populations genetically as it is to say that we can distinguish between the colors red, blue, and yellow. The take-home message, though, is that one should be extremely careful about drawing conclusions that extend beyond the populations that have actually been sampled—in this case, the genetic differences among the three HapMap populations do not tell us how genetic variation is actually organized across the entire human species.

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

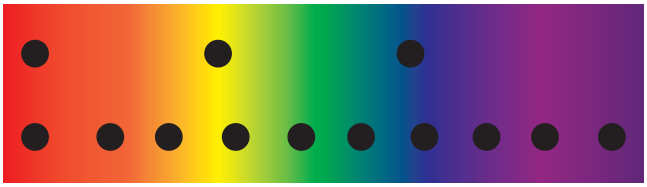


FIGURE 8.1

Different views of a rainbow, depending on how one "samples" the rainbow (black dots). The top part shows that if just three parts of a rainbow are sampled, one might conclude that a rainbow consists of very different components, namely, red, yellow, and blue. The bottom part shows that if one instead samples many different parts of the rainbow, one would instead conclude that a rainbow is a gradient of colors with no discrete boundaries between them.

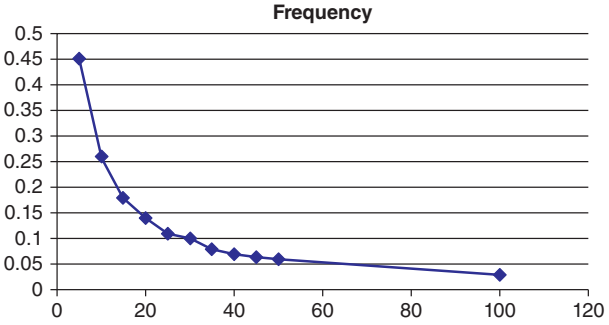
As you might imagine, this analogy extends to other sampling scales. Suppose I want to study how genetic variation is distributed among the major language families of the world, such as Indo-European, Austronesian, Niger–Congo, and so forth. Would it be sufficient to sample one group from each language family? Maybe it would be—but the only way to tell would be to sample multiple groups from each language family, in order to see how the variation among groups from the same language family compares to the variation among groups from different language families. One should be extremely cautious about studies that draw broad or sweeping conclusions based on sampling just one group from a region, language group, ethnic group, or larger population group of interest. The extent to which such results can be extrapolated beyond the actual groups sampled depends heavily on assumptions as to how "representative" the sampled groups are of the larger entities of interest, and the only way to be really certain about this is to actually have the data from several groups from each entity of interest.

Even with a plan as to which groups to sample in order to investigate the questions of interest about population history, there are other practical issues to consider. All analyses of genetic data start with the assumption that individuals have been sampled randomly, but it is not clear what exactly that means. Literally, a random sample would mean that individuals have been sampled from a location without any regard to any other characteristics, including sex, age, ancestry, languages spoken, relatedness with others, and so forth. In practice, we have to sample individuals old enough to give informed consent (more about this later), we often prefer to sample males so as to study Y chromosome variation, and we often try to restrict the sampling to individuals with ancestry from the geographic region and/or language group of interest by asking sample donors for information about where the parents (and, if known, grandparents) were born and/or what language(s) they spoke. Moreover, we often try to avoid sampling close relatives (siblings, parents and offspring, etc.) because they will share many genes in common, and it seems a waste of time and resources to carry out genotyping for close relatives when we can predict what the results will be. But it might actually be the case that a truly random sample would differ in important characteristics from sampling under the above conditions—for example, if a particular group does in fact contain a relatively high proportion of related individuals, then by focusing the sampling on only unrelated individuals, we might end up with a biased view of the actual genetic diversity in the population. There are no easy answers to the question as to how to choose individuals for inclusion (or exclusion) in a study, but what can (and should) be done is to carefully document (and publish) the criteria that were used, so any effect of the sampling design on subsequent analyses can be readily ascertained.

Another practical question of interest is the number of individuals that need to be sampled. The flippant answer is "as many as possible," but the reality is that one has a limited amount of time in the field and a limited amount of resources to devote to the genotyping, so one has to set limits. To some extent, the desirable sample size depends on the questions of interest and the types of analyses to be done. If you are interested in the distribution of very rare alleles, for example, then you may need sample sizes in the hundreds or thousands. On the other hand, if you are intending to carry out complete genome sequencing, then even one individual can suffice for many analyses (e.g., Figure 12.19)—as long as that one individual is truly "representative" of the group of interest. But for the "typical" molecular anthropology study, which might assay various genetic markers on mtDNA, the Y chromosome, and/or the autosomes, a general rule of thumb is that sample sizes should be in the range of 20–40 chromosomes (i.e., 10–20 unrelated people for analyzing the diploid autosomes, and 20–40 unrelated people for analyzing the haploid mtDNA or Y chromosome)—fewer than 20 runs the risk that the genetic variation in the population is not adequately sampled, while the amount of information gained from more than 40 is generally not worth the extra effort (although keep in mind that larger sample sizes are nonetheless desirable for many analyses, such as assessing the existence of subgroups, or if one is specifically interested in rare alleles). In case you are interested where this general rule of thumb comes from, Box 8.1 gives some additional details. However, in some cases, especially where the population is of particular interest and it simply wasn't possible to obtain more samples, sample

BOX 8.1 ■ Deciding How Many Samples to Collect

The number of samples one needs for a study depends on the question(s) of interest and the analyses to be undertaken, so there is no one general rule that holds for all situations. But for the "typical" molecular anthropology study, we want to have reasonably accurate estimates of the genetic diversity within populations and the genetic differences among populations (how we actually measure these is discussed in detail in Chapter 10). This, in turn, means that we need to have sampled the most frequent alleles in the populations—rare alleles don't contribute as much to estimates of genetic diversity or genetic differences. So, one way to approach the sample size question is to ask how big a sample size do we need to have a 95% chance or greater of sampling an allele at a particular frequency? The figure shows precisely that; the sample size is on the X-axis, while the Y-axis shows the lowest frequency of an allele in the population that you can expect (with 95% probability) to detect in that sample. For example, with a sample size of 20 you can expect to detect an allele with a frequency of at least 14% in the population, while with a sample size of 40 you would detect an allele with a frequency of at least 7%. Notice that increasing the sample size beyond 40 or so doesn't gain you much; with a sample size of 100 (so, sampling 60 additional individuals) you can expect to detect only an allele with a frequency of at least 3%. So, sample sizes of 20–40 are sufficient to ensure that you detect the alleles that are most important when it comes to estimating genetic diversity and genetic differences, but do keep in mind that other analyses may have much different sample size requirements.



sizes as low as 10 chromosomes (or even lower) are in fact used.

SAMPLING POPULATIONS: ETHICAL ISSUES

A full discussion of the various ethical issues that need to be considered when obtaining biological samples from humans is beyond the scope of this book; however, a brief mention of the more important points, especially those that pertain to obtaining samples for studies of human genetic history, is warranted. In this section, we will discuss issues related to sampling of living individuals for a new study, while in the next section we will consider the use of archival samples (i.e., samples already collected, often for another purpose, from donors who may be dead or otherwise unavailable for further contact). First, all research institutions have institutional review boards (IRBs) to review and grant ethical approval for research dealing with human subjects; moreover, most countries require research permits in order to obtain and take away human biological samples. It should go without saying that investigators must comply with such requirements, and yet we are occasionally contacted by well-meaning but uninformed people who are in, or planning to soon go to, a location where they think it would be interesting to collect samples and ask whether we would like them to do so. Unfortunately, obtaining IRB approval and research permits usually takes weeks—if not months—and, therefore, requires planning in advance, so such spontaneous requests can seldom be accommodated.

Second, in addition to institutional IRB approval and government research permits, in some areas of the world it is necessary to also obtain the approval of a tribal or indigenous people's organization, chief or headman, or other entity that is responsible for overseeing the welfare of particular groups. But most importantly, even with all required IRB approval, research permits, and permission from other responsible entities, it is still necessary to obtain free, prior, and informed consent (sometimes abbreviated FPIC) from every individual who donates a sample. "Free" means that the persons give the consent of their own free will, without coercion, and also that they are capable of deciding for themselves if they want to participate or not (e.g., we avoid taking samples from children or from people who are obviously inebriated). "Prior" means that prospective donors are told what will be done with their samples before the samples are collected, so that they can then refuse to participate if they like. And "informed" means that the prospective donors are given sufficient information about the study, and all questions answered to their satisfaction, so they can decide whether they want to participate or not. This last part is perhaps the most troublesome, as how does one explain the goals of a molecular anthropology study such that individuals with limited educational background and/or limited knowledge of genetics can really understand what will be done with their samples? Some ethicists have

BOX 8.2 ■ Explaining to Prospective Sample Donors Why We Study Genetic History

Prospective sample donors come from a wide variety of backgrounds and have varying degrees of knowledge, so the explanation has to be tailored to the individual situation, as well as the goals of the particular project (e.g., if we want to compare the genetic relationships of groups speaking different languages, then we might include some discussion of the languages). But in general, we start by introducing who we are and where we come from, and that we have come to them because it is our job to learn about the history of people from different parts of the world. We then explain that there are many different ways to learn about history—from reading books, from talking to the elders, from digging up things from the ground, and so forth—but our job is to learn about history from genes. This is followed by a brief and nontechnical description of what genes are (e.g., genes are small things that you cannot see inside your body, and they are the instructions that tell the body how to grow and develop). We then explain that you get your genes from your parents, who in turn got their genes from their parents, who in turn got their genes from their parents, and so forth, so everybody today has genes that came from people who lived a long time ago. This then helps make it clear how one can learn about people from a long time ago by studying the genes in people today.

(in the context of disease-related studies) gone so far as to advocate that after the goals of the study are explained to the prospective donors, they should have to pass an examination that tests their comprehension before being permitted to donate a sample! This, in my view, is overkill; in my experience, if you take the time to explain carefully and patiently what it is you want to learn from doing this work and answer all questions (either directly in the lingua franca or with the aid of a knowledgeable translator), then people are perfectly capable of deciding for themselves if they think what you want to do is something they approve of or not—and it certainly does happen, after spending hours in explanation and answering questions in front of a group of people, that nobody offers to donate a sample! For those of you who are curious as to what we actually tell prospective donors, Box 8.2 gives more details.

Along with explaining the goals of the study, it is important to let the prospective donors know about any risks they may incur by participating. Such risks include potential physical injury or discomfort, as well as negative consequences that might arise from publication of genetic data—the latter is of more concern for disease-related studies, where, for example, the fear is that if it becomes known that an individual (or a population) has a higher genetic susceptibility for a disease, this information might be used to deny the individual/population insurance coverage or lead to other discrimination. Nowadays, we routinely collect saliva for our studies, which involves simply having the donor spit into a tube that contains a harmless buffer to preserve the DNA. So, beyond a dry mouth there is no discomfort for the donor, which is a big improvement over sticking needles into people to collect blood! Sample donors are also assured that their samples are anonymized, as only a sample ID number goes on the tube itself, no names or other identifying information. All such identifying information (name, age, sex, birthplace, language(s) spoken, etc.) is recorded (along with the sample ID number) on information sheets bound into a sampling book, which the donor signs (or marks, for donors who are illiterate) to indicate his or her consent to the use of his or her sample for the study. The sampling book is then kept in a secure location, and no identifying information is published with the genetic data; thus, the published genetic data cannot be linked to specific (named) individuals—except, as discussed below, it may (with a lot of effort) nonetheless be possible to identify who donated a particular sample, just from publicly available information.

However, publication of genetic data raises other ethical concerns. In general, data from scientific studies should be made available to the entire scientific community, and doing so is strongly encouraged, if not absolutely required, when a study is published. The benefits of making genetic data publicly available are readily apparent when one sees the enormous use that has been made of the public repositories described in the next chapter. But with the increasing use of genome-wide data (either from SNP chips or genomic sequencing) in molecular anthropology studies, there is the risk that putting such data in a public repository will lead to others using the data for purposes that contravene the informed consent. For example, other investigators might search the data for genetic variants associated with susceptibility to particular diseases, even though the research permit and/or informed consent forms expressly prohibit any disease-related research with the samples. The current solution—which is admittedly imperfect, but seems the best that can be done—is to restrict access to such genetic data to investigators who promise to adhere to the restrictions under which the samples were collected, for example, the data can be used only to address questions about population history and not about disease.

The same solution holds for attempts to identify the individual donors from the genetic information. Recent studies have made headlines by showing that it can be surprisingly easy to take an individual's genetic data from a public repository and with a bit of detective work come up with a very good guess as to the identity of the supposedly anonymous donor (Gymrek et al. 2013). In case you are wondering how that could be possible, one way is to take advantage of the widespread "community" databases of Y chromosome types that are linked with surnames. Recall that the Y chromosome is present only in males and transmitted from fathers to sons, so the expectation is that Y chromosome variation should be strongly associated with surnames (at least, in those cultures where the son takes the surname of the father), and indeed that does seem to be the case—with the inevitable exceptions due to nonpaternity, adoptions, and the like. Thanks to the rising interest in personal ancestry, you can pay a fee and have your Y chromosome typed (more about methods for determining Y chromosome variation in the next chapter) by a company and then go online to see who else has similar Y chromosome types—sometimes you can even identify long-lost relatives this way. So to identify an anonymous donor in a public database, what can be done is to take the Y chromosome information in the database for an individual (for males, of course!), search the public community Y chromosome databases for matching types, and now you have an associated surname. This, along with the age and geographic location of the donor (also usually provided in the database), is often enough to identify the anonymous donor, especially with the growing amount of information available via social media. Some researchers have even argued that since it is futile to guarantee the anonymity of donors to genetic databases, all information associated with a sample in a database should be made available—even medical records! But in my view (and that of many), decisions about what information to make public and what to keep private are up to each individual to decide—nobody should feel pressured to make any information public that they would prefer to keep private. So, the way we handle concerns about attempts to identify sample donors is to require anyone gaining access to the genetic data we generate to promise not to make any such attempt—maybe not the best solution possible but one that works while still making it possible for responsible investigators to have access to the data.

Another important ethical aspect of sampling for molecular anthropology studies has to do with the return of benefits to the individuals or the community. It is usually not permitted to pay donors for their samples, as this might coerce people to donate a sample who would otherwise be unwilling to participate—although small gifts in recognition of the time it takes people to listen to the project explanation and donate a sample are OK, so, for example, we've given donors small bags of tea and sugar in Namibia, or a bit of fishing line and fishhooks in the Solomon Islands. Otherwise, what molecular anthropologists can return to the community is what they themselves gain from the work: namely, knowledge about the genetic history of the community. This return of knowledge can take many forms. Some researchers remain in close contact with communities and thereby regularly communicate results—the research can even be shaped by the ongoing dialogue between the researchers and the community. For more remote communities, such regular, ongoing contact may be impossible, but it should always be possible to communicate the results back to the communities after the research is done. A follow-up visit by the researchers (or someone who can knowledgeably explain the results—for example, linguists or social anthropologists working with the communities) would be best but may not be practical due to budgetary or other constraints. In such situations, one can then send posters or flyers that explain the results in nontechnical terms—and translated into the local lingua franca—to central locations (e.g., regional government offices) to be distributed more widely to schools, clinics, village/tribal offices, and so forth.

While some researchers provide genetic results to each individual donor—indeed, this can be a way of attracting interest in participating in such studies—as a general policy, we do not inform individuals about their own genetic results, as there is always the possibility that individuals will learn something unexpected or disconcerting about their own personal genetic history. Instead, we provide results about the genetic history of the community. But even when communicating the genetic results to communities, it is important to be sensitive to issues such as cultural identity (e.g., the genetic results may indicate that the group's genetic relationships do not match their cultural identity), land rights (e.g., the community may be involved in a dispute over land rights with another group and, therefore, be keen to use genetic results that show that they were in that area first to support their claims), and so forth. In our experience, the best way to deal with such issues is to emphasize that genetic history provides only one view of history and not a very important one at that when it comes to thinking about identity—after all, what defines you as a person is a lot more than just the genes you inherited from your ancestors via your parents: it's all of your experiences, the language(s) you speak, interactions with your peers, your educational background, and so forth. The same with land rights—genetic history gives us only some limited insights into events that happened thousands and thousands of years ago, which are purely of academic

interest and should not have any bearing on how communities today resolve such disputes.

There are other ways to transfer knowledge beyond simply communicating the results of the study. While in the field, one can offer to teach some basic genetics at local schools or give public lectures—one of my fondest fieldwork memories is of a public lecture on the peopling of the Pacific that I gave in Gizo in the Solomon Islands, out in the open under the stars before a rapt audience, with only an inflatable globe for a visual aid! Local scientists can be invited to your institution for training in laboratory methods and/or data analysis, thereby spreading such knowledge to their colleagues and students. Local students can also come for such training, either as interns or even as graduate students—such a PhD degree can greatly enhance the individual's career opportunities in their home country and help strengthen scientific research in countries with limited resources for such work.

Finally, there are other steps that can be taken to provide additional benefits to the places we go to in order to collect the samples that allow us to carry out our work and advance our careers. For example, one can bring: hard-to-obtain chemicals and books for local scientists; medical supplies for local clinics—even something as simple as vitamin supplements; and/or supplies such as pencils and paper for local schools. It can be a sobering experience to see the extraordinary circumstances that scientists, doctors, and teachers labor under in some parts of the world, and a little assistance can go a long ways.

ARCHIVAL SAMPLES

The previous section discussed issues related to current sampling of populations, but archival samples (i.e., samples collected for other studies, sometimes decades or even centuries ago) are another important source of samples for molecular anthropology studies and raise some different issues. On the one hand, such samples can be extremely valuable, if not irreplaceable—archival samples exist for some groups that do not exist as such anymore, because the groups either have literally become extinct or (more often) have dispersed and/or become amalgamated with other groups. Even if the groups still do exist, it may be much easier (and more justifiable from both a scientific and ethical standpoint) to work with archival samples than to go to the time, trouble, and expense of mounting an expedition to try to collect samples that are already available.

On the other hand, the use of archival samples raises the troubling issue as to whether or not the appropriate consent was obtained for the use of the samples in molecular anthropology studies. If some sort of consent was obtained initially for studies related to population history, then in general there is no problem with using them for molecular anthropology studies (even though the methods may have changed considerably), as long as the usual standards of anonymization of samples are followed in order to protect individual privacy. It must be kept in mind that informed consent standards differed in the past; frequently only oral consent was obtained, so whether or not the donors actually gave informed consent for their samples to be used for population history studies relies on the memory (and honesty) of the people who collected the samples.

When specific consent for population history studies is lacking, then the use of archival samples for molecular anthropology studies becomes much murkier. Some would argue that as long as the samples are either anonymous or anonymized, then they can of course be used for molecular anthropology studies—after all, there are clear scientific benefits and no harm done to the sample donors, so what's the problem? Others, however, take the view that any use of the samples that is not expressly permitted is prohibited. In particular, if consent was obtained only for disease-related research, then perhaps such samples should not be used for population history studies, because while the donors agreed to the disease-related research, there is no evidence that they would have agreed to donate samples for population history studies—maybe they would not want their samples used for such purposes. And for those of you who think anything not prohibited should be permitted, there is the sobering example of the Havasupai, a native American tribe from the southwestern United States who in the 1990s donated blood samples to researchers from Arizona State University. The samples were donated for research into genetic mutations that might lead to diabetes, which occurs at an extraordinarily high frequency in the Havasupai. But the samples were also used for research into other medical conditions, including mental illness, as well as for studies of population history. When the Havasupai found out about these additional, unauthorized studies, they sued—and despite the claims of the Arizona State researchers that the Havasupai had given broad consent for genetic studies, the university settled with the Havasupai, giving them $700,000 and returning their DNA samples to them. Moreover, this is not the only example where communities have asked for DNA and/or blood samples to be returned to them. While it seems clear enough that any individual who donated a sample for research has the right to change his or her mind and ask that his or her sample not be used for research, in some cases it is a community or an organization and not the actual sample donors who ask for research to cease and/or samples to be returned. How to balance the scientific benefits to be gained from research with such samples against the desires of the

community asking for their return, especially in cases where the community asking for their return has no demonstrable close relationship to the sample donors, remains a difficult issue.

Another situation for which there are no clear-cut answers arises with samples for which no consent was obtained. It should go without saying that samples taken by force or under coercion should not be used for research, regardless of the potential benefits. But what if there is no evidence one way or another concerning consent? For example, while I was a graduate student in the early 1980s, I received some placentas from aboriginal Australians, collected from various hospitals by an Australian researcher, which I used in my research. What type of consent—if any—had been obtained was not something I thought about, as in the early 1980s informed consent was not the issue it is today, plus that's the sort of thing for advisors, not students, to worry about (or so I thought). These samples were used in my PhD research and in several subsequent studies. It later came out that the Australian researcher may have simply taken the placentas without asking the individuals for their consent—not so surprising, because then (as now) placentas were routinely disposed of after hospital births without anyone asking the parents whether they cared about what was done with their placenta. So, some would argue that no consent is needed to use placentas for research since they are otherwise just going to be destroyed. But again, if the individuals had been asked, maybe they would have said that no, they didn't want their placentas used in studies of population history. Because there is no consensus on this issue, I have stopped using these samples in my research.

Currently, how to deal with such archival samples remains a thorny issue. Research on living human subjects is governed by the principles of the Declaration of Helsinki; what is desperately needed is a similar set of guidelines for using archival samples that takes into account both the scientific value of such research and the ethical responsibilities underlying the use of such samples. In the meantime, one way forward is to contact the communities from which the archival samples originate, or organizations that oversee research involving these communities, and seek their approval for the research. A nice example is provided by the full genome sequence of an aboriginal Australian that was determined from a hair sample obtained by the British ethnologist Alfred Cort Haddon in 1920 as he traveled across Australia. Ethical concerns over the use of the hair sample for genome sequencing were raised when the scientists tried to publish their work, so the principal investigator, Eske Willerslev from Denmark, went to Australia and contacted the tribal council that represents communities in the general geographic area from which the hair sample was collected and was successful in gaining their permission to publish (Rasmussen et al. 2011). This sort of compromise between having the free, prior and informed consent of the actual individual involved versus no consent whatsoever is probably the best that can be achieved in such instances.

Similar concerns arise with samples for which no consent was ever possible, for example, analysis of ancient DNA from skeletal remains that are hundreds to thousands of years old. While it used to be thought that such skeletal remains are fair game for any sort of scientific research, requests for repatriation of such remains are on the rise, sometimes from groups that have no demonstrable connection to the remains. In sum, there is a clear need for a consensus set of guidelines on when and how archival samples should (and should not) be used for molecular anthropology studies.

CHAPTER 9

SAMPLING DNA REGIONS

We've seen in the previous chapter some issues that arise with sampling individuals and populations; the choice of which region(s) of DNA to study can also have a major impact on the results of a molecular anthropology study. We've already discussed some properties of different DNA regions that clearly illustrate this, in particular how fast a particular DNA region evolves. For example, if you want to investigate the genetic relationships of some human groups that are likely to be quite closely related, then you probably should not choose to investigate histone genes, since (as we saw in Chapter 6) histone genes evolve so slowly that you wouldn't detect any genetic variation. Conversely, minisatellites would be a poor choice for investigating the relationships among humans, chimpanzees, and gorillas, because minisatellites evolve so rapidly that one could not distinguish which pair of these three species are most closely related. So it is important to follow the "Goldilocks" principle when designing a study—that is, you don't want to use genetic markers that aren't variable enough, or are too variable, but instead are "just right" for the questions you are interested in.

But there is more to it than just choosing genetic markers with enough—but not too much—variability; it turns out that humans (like other creatures!) are blessed with some different types of DNA that are well suited for particular kinds of questions, and we will now discuss these and the properties that make them useful for molecular anthropology studies.

MITOCHONDRIAL DNA

For many decades after the discovery of DNA, it was thought that all of your DNA resides as chromosomes in the nucleus of the cell. It, therefore, came as quite a surprise in the early 1960s when it was discovered that mitochondria, which are small structures or **organelles** in the cytoplasm of the cell (Figure 9.1) that are involved in energy production, have their own DNA (Nass and Nass 1963a, 1963b; Schatz et al. 1964). Moreover, it turns out that mitochondrial DNA (**mtDNA**) has several rather peculiar properties that distinguish it from the chromosomal DNA in the nucleus.

First, as shown in Figure 9.2, mtDNA is a circular molecule (as opposed to the linear chromosomes in the nucleus) and is quite compact. It has only about 16,500 bases (compared to about 3.2 billion bases in the haploid nuclear genome) and contains just 37 genes: 13 protein-coding genes, 2 ribosomal RNA (rRNA) genes, and 22 transfer RNA (tRNA) genes. The 13 protein-coding genes include three subunits of cytochrome oxidase, two subunits of F_1-ATPase, seven subunits of NADH–dehydrogenase, and the gene for cytochrome B. All of these are involved in the main function of mitochondria, which is to carry out cellular respiration, which in turn involves the production of energy from metabolites. All of the polypeptides encoded by the mtDNA genes combine with other polypeptides encoded by nuclear DNA to form the active protein complexes involved in the production of energy. Moreover, all of the proteins needed to replicate mtDNA and to transcribe, process, and translate messenger RNA (mRNA) from mtDNA are also encoded by the nuclear DNA, and hence all of these nuclear-encoded mitochondrial proteins must be imported into the mitochondria.

Thus, cells have to produce several hundred proteins to maintain mtDNA and allow it to function, and in return mtDNA provides just 13 protein subunits, which hardly seems like a fair trade. This raises the question as to where mtDNA came from in the first place and why we still have it. It turns out that mtDNA is a relic of an **endosymbiosis** event that occurred more than a billion years ago: a primordial bacterial (or bacterial-like) cell merged with another cell and gradually the two cells adapted to coexist with one another, with one cell evolving to specialize in energy

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

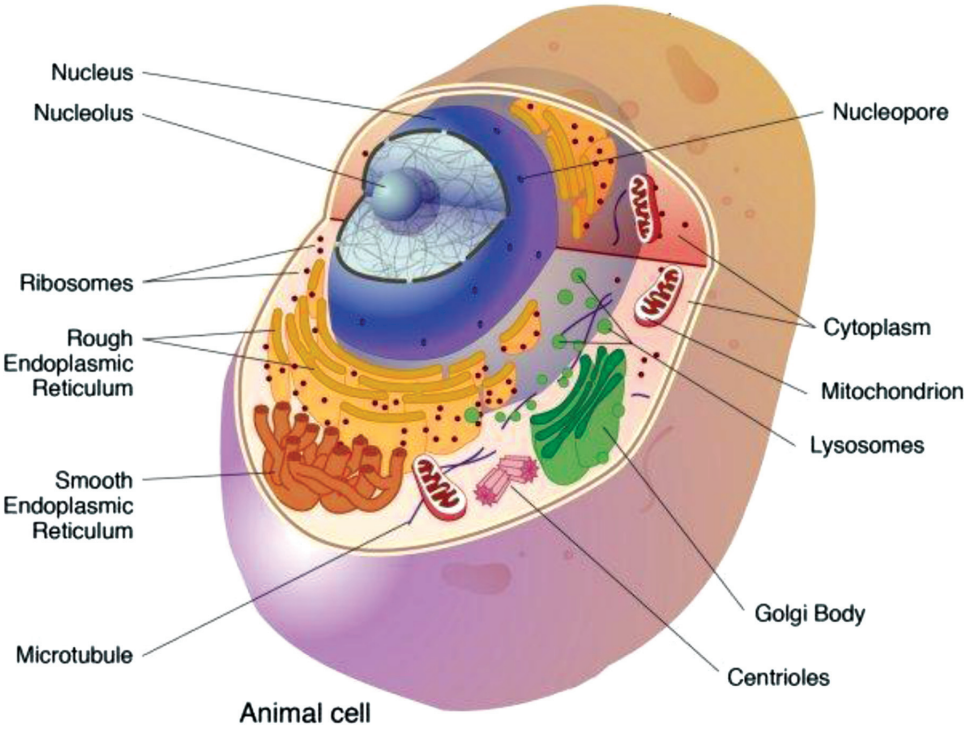


FIGURE 9.1

A typical animal cell, showing the various structures. Two of these structures, the nucleus and the mitochondria, contain DNA. Reprinted with permission from the National Human Genome Research Institute (http://www.genome.gov/Glossary/resources/cell.pdf).

production and hence becoming the mitochondria. Sounds pretty crazy—and that's exactly what people called the late Lynn Margulis when in 1967 (writing as Lynn Sagan, as she had been married to Carl Sagan) she proposed that mitochondria and chloroplasts both had their origins via endosymbiosis (Sagan 1967), and moreover that these were critical events in the evolution of animals and plants (because mitochondria convert oxygen to the energy necessary for us, multicellular organisms, to survive, while chloroplasts allowed plants to proliferate by converting carbon dioxide to oxygen—which in turn fueled animal evolution). By her own account she submitted her seminal paper outlining her theory of endosymbiosis to more than a dozen scientific journals before it was accepted by one for publication. Margulis persevered in sticking to her ideas despite years of ridicule and criticism and was rewarded for her tenacity by seeing her ideas gain widespread acceptance when the first genome sequences of mtDNA and chloroplast DNA revealed that they were not merely extensions of nuclear DNA but indeed had a separate origin. Current thinking is that following the endosymbiosis event that led to the origin of mitochondria, as the progenitor cell for mitochondria evolved to become more and more specialized for energy production, the genes for other functions were transferred to the nucleus. There is plenty of evidence for transfer of DNA sequences from the mitochondria to the nucleus over evolutionary time—there are more than 1000 mtDNA-related sequences that can be identified in the human nuclear genome, and some of these **NUMTs** (nuclear copies of mtDNA sequences) have transferred to the human nuclear genome so recently that polymorphism exists for the presence or absence of a particular NUMT at a specific location in the genome. So, over evolutionary time mtDNA sequences have been transferred to the nucleus and deleted from the ancestral mtDNA genome, and mtDNA sequences are still ending up in the nuclear genome even today.

So why do we still have mtDNA—surely it would be a lot more efficient if the mere 13 protein subunits still encoded by mtDNA were also transferred to the nuclear DNA, and thus the cell would no longer need to make the hundreds of proteins needed for mtDNA maintenance and function? The answer is that we don't know for sure, but it probably has something to do with another peculiar (and astonishing!) property of mtDNA: namely, mtDNA has a different genetic code than nuclear DNA (recall that the genetic code, discussed in Chapter 2, is the correspondence between codons in DNA and amino acids in polypeptides). Shortly after the rules of the genetic code were worked out in the 1960s, the genetic code was determined to

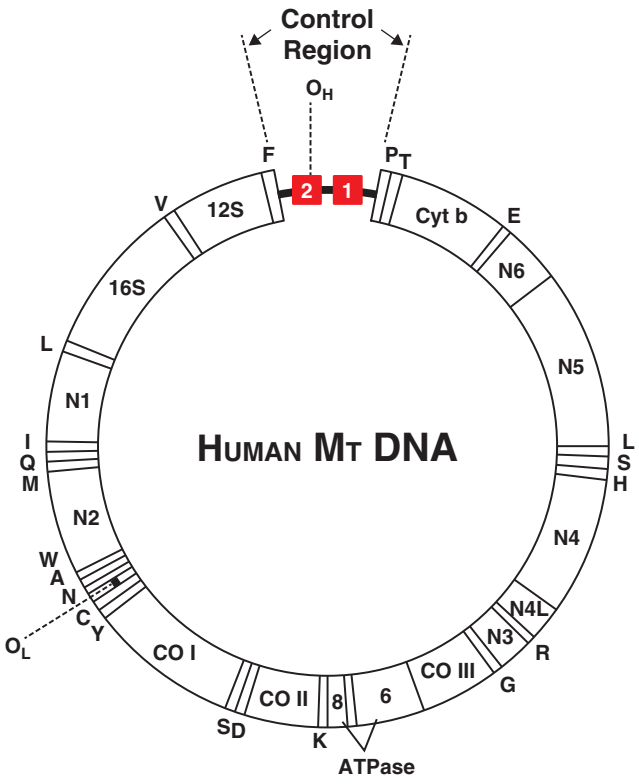


FIGURE 9.2

Circular human mtDNA molecule. There are 13 protein-coding genes (N1-N6+N4L, COI-COIII, ATPase 6 and 8, and Cyt b), two rRNA genes (12S and 16S), and 22 tRNA genes (indicated by single letter code). There are two origins of replication, one for each strand (O_H and O_L), and most of the noncoding DNA is found in the control region, which contains two hypervariable segments denoted HV1 and HV2 (indicated by the numbered red boxes).

be the same in bacteria, yeast, plants, fruit flies, mice, humans—in short, in every living thing examined—so it was assumed that the genetic code was universal. It was, therefore, truly shocking when it was discovered that mtDNA uses a different genetic code—and even more shocking when it was discovered that the mtDNA genetic code varies among organisms (reviewed in Fox 1987)! In vertebrates there are four differences between the mtDNA genetic code and the (so-called) universal genetic code (UGA is not a stop codon in mtDNA but encodes the amino acid tryptophan; AGA and AGG are stop codons in mtDNA instead of encoding arginine, and AUA encodes methionine in mtDNA instead of isoleucine). Fruit flies and other invertebrates, yeast, ciliates, flatworms, echinoderms—all of these groups (and more) each have their own mtDNA genetic code that differs from the universal genetic code and from each other.

How is it that mtDNA has a different genetic code that even varies among organisms? The answer—again—is that we don't know for sure, but one possibility is that back in the distant past a mutation arose in some tRNA gene in the mtDNA that caused a change in the genetic code, and so codons became reassigned via a process known as codon capture (Osawa and Jukes 1989). Ordinarily, we would expect such a change to be lethal—after all, if you arbitrarily substitute one amino acid for another across hundreds or thousands of proteins, for sure you're going to disrupt the function of at least a few of them, with dire consequences. However, perhaps when there were only 13 or so protein subunits being made by mtDNA, this shift in the mtDNA genetic code wasn't lethal, and the mtDNA protein-coding genes subsequently evolved to adapt to the code shift. But once they had adapted sufficiently to the new genetic code, they would no longer function with the old genetic code and hence would not be able to function in the nucleus. According to this view, mtDNA is thus a "frozen accident" (a term first coined by Francis Crick of Watson and Crick fame to describe the origin of the genetic code itself; Crick 1968) that we are stuck with because of this genetic code shift. If this is correct, it hardly supports the view of any "intelligent design" when it comes to mtDNA!

Another peculiar property of mtDNA is that there is very little noncoding DNA: only about 7% of the mtDNA genome is noncoding DNA, compared to about 98.5% for nuclear DNA. Moreover, mtDNA genes are very compact, as there are no introns (noncoding DNA within a gene) and very little noncoding DNA between genes—where one gene ends, another immediately starts. In fact, some genes even have an incomplete termination codon and require polyadenylation (addition of adenine nucleotides to the end of the mRNA) to form a complete termination codon—a rather extreme example of evolution getting rid of as much nonessential DNA as possible! The noncoding DNA occurs mostly in the so-called control region (Figure 9.2), which varies in length among species (it's about 1100 bases in humans) and gets its name from the fact that it contains one origin of replication and both origins of transcription, as well as functional elements involved in initiating and terminating replication and transcription. That is, each strand of the mtDNA double helix has a single place where DNA replication starts and a single place where mRNA transcription starts; the entire mtDNA genome is transcribed from each strand into a single RNA molecule, which is then processed to make the individual mRNA, tRNA, and rRNA transcripts.

This extreme compactness of the mtDNA genome suggests that it has been subject to strong evolutionary pressures to be as small as possible and still maintain necessary functionality. Accordingly, this selection

to prune away anything nonessential in the mtDNA genome should also be manifest at the sequence evolution level, and hence it was expected that mtDNA would have a slow evolutionary rate. Thus, it was quite surprising when it was discovered that mtDNA actually evolves some 10 times or so *faster* than nuclear DNA (Brown et al. 1979). Why this should be the case is not known for sure, but we do have some pretty good ideas. As was discussed in Chapter 2, new mutations can arise either because of mistakes during DNA replication or because of DNA damage that is not repaired accurately. MtDNA replication is carried out by a different set of enzymes than those used in replicating nuclear DNA, and initially it was thought that the mitochondrial DNA polymerase might have a higher error rate than the nuclear DNA polymerases. However, subsequent work has not demonstrated a higher error rate for the mitochondrial DNA polymerase (Lee and Johnson 2006). Instead, attention has focused on the repair of damage to mtDNA. As a by-product of metabolism, several substances are produced in the mitochondria that are potent DNA-damaging agents, so there is plenty of potential for DNA damage to occur at a higher rate in mtDNA. Moreover, it was initially thought that there was no repair of damage to mtDNA; however, subsequent work as shown that there is lots of repair going on in mtDNA, and it remains to be seen whether there are any deficiencies in the repair of damage to mtDNA that could explain the high mutation rate (Alexeyev et al. 2013). Anyway, regardless of the underlying reason for the higher rate of mtDNA evolution, it turns out to be an extremely useful property for human population history studies, as you get more bang for your buck from analyzing mtDNA than from nuclear DNA (i.e., many more polymorphisms from sequencing the same amount of DNA). So, in accordance with the aforementioned Goldilocks principle, mtDNA variation is "just right" when it comes to studying human population history.

Another useful property of mtDNA is that it is present in multiple copies per cell—the average mitochondrion has 5–10 mtDNA genomes, and the average cell has a few hundred to a few thousand mitochondria—compared to just two copies of any nuclear DNA gene. This greater abundance of mtDNA than nuclear DNA, plus the localization of mtDNA in a cytoplasmic organelle separate from the nucleus, initially (before the development of the polymerase chain reaction) made it relatively easy to isolate and analyze mtDNA (i.e., if you consider it relatively easy to prepare and carry out cesium chloride density gradient ultracentrifugation from tissue extracts for 2 weeks to get the mtDNA from one placenta, which is what we did when I was a graduate student!). It also makes mtDNA the genome of choice for analyzing DNA from ancient specimens (as discussed in more detail later), as well as from certain kinds of forensic specimens (old bones, hairs, burnt remains, etc.) because in any specimen where there is likely to be very little (if any) DNA surviving, the fact that mtDNA exists in many more copies than nuclear DNA greatly enhances the chance of success.

The final property of mtDNA that makes it of interest to molecular anthropologists—and perhaps the most extraordinary one—is that mtDNA is maternally inherited. Both males and females have mtDNA, but you get all of your mtDNA from your mother and none from your father. Maternal inheritance of mtDNA was demonstrated beginning in the 1970s (Giles et al. 1980; Hutchison et al. 1974) and at first was thought to be due to exclusion of paternal mtDNA from the egg during fertilization. Sperm do have mtDNA—indeed, the midpiece of the sperm, directly behind the head, contains some 50–100 mitochondria that provide the energy for the sperm to swim—but it was initially thought that only the head of the sperm, which doesn't have any mitochondria, entered the egg upon fertilization. However, subsequent studies showed that the sperm midpiece does enter the egg (Ankel-Simons and Cummins 1996).

The next idea about maternal inheritance of mtDNA was that perhaps it is simply a matter of numbers. The maternal mtDNA is amplified to upward of one million copies shortly before fertilization, apparently because there is no mtDNA replication during the first few rounds of cell division following fertilization, so the mtDNA in the fertilized egg has to be partitioned among the increasing numbers of cells. With one million maternal mtDNA copies versus (at most) a few hundred paternal mtDNA copies, maybe the paternal contribution is simply undetectable in the face of the massive maternal contribution. To test this hypothesis, studies were undertaken in fruit flies and in mice that attempted to enrich for any paternal mtDNA contribution (Gyllensten et al. 1991; Kondo et al. 1990). This was done by crossing females from one species to males from a second species, taking the female offspring and crossing them back to males from the second species and repeating this process for many generations. The idea was to build up a detectable level of paternal mtDNA from the second species, and using females and males from different species was done so that their mtDNAs could be readily distinguished. And these experiments found that the paternal mtDNA could indeed be detected after such enrichment. Unfortunately, when you do interspecies crosses in the laboratory, you often get bizarre results that don't reflect at all what happens in nature, and that seems to also have been the case here: when these experiments were repeated using females and males from different strains within the same species of mouse, no paternal mtDNA could be detected; further experiments showed that

within a mouse species, shortly after fertilization the paternal mitochondria are segregated and destroyed, while apparently in the interspecies crosses the mechanism by which this happens breaks down (Kaneda et al. 1995).

Still, the interspecies results showed that the mechanism responsible for maternal mtDNA inheritance is under genetic control, and like any such genetic process the possibility always exists for mutations to occur that disrupt the mechanism. And indeed, in 2002 a case of paternal inheritance of mtDNA in humans was reported (Schwartz and Vissing 2002) in a family that was being studied because of a disease associated with mtDNA. This study is noteworthy in that extensive control experiments were conducted to rule out any possibility of contamination or some PCR artifact as the explanation for the results. While one can never be certain that some unknown experimental artifact is actually responsible, currently the best explanation for the results does appear to be paternal mtDNA inheritance—although one should keep in mind that the paternal mtDNA in this individual carried a novel deletion in one of the mtDNA coding genes, which likely rendered it nonfunctional. Nonetheless, after decades of studies involving thousands of families, this remains the single example of paternal mtDNA inheritance in humans—all other studies have found, without exception, only maternal mtDNA inheritance. Whether paternal mtDNA inheritance is truly as rare as current results suggest (and may be limited to mtDNA with peculiar defects), or whether there actually is more low-level paternal inheritance than can be detected with current methods, will be shortly resolved (probably by the time this book is published!) with ongoing large-scale, next-generation sequencing studies of families.

In the meantime, in this book we take the view that mtDNA is, for all practical purposes, strictly maternally inherited with no recombination. Even if mtDNA did undergo recombination, the fact that all of your mtDNA genomes are identical (or nearly so, as there can be **somatic mutations** occurring in the mtDNA of some of your cells as you develop and age) means that recombination simply swaps identical segments among mtDNA genomes and hence has no detectable effect. This strict maternal inheritance with no recombination has two important consequences. First, mtDNA provides insights into the maternal history of populations. Many aspects of humans and their societies are sex-biased (i.e., involve or influence males and females differently) and can also have an impact on patterns of genetic variation; the comparison of mtDNA with Y chromosome (discussed in the next section) variation can thus be particularly informative when investigating such aspects, and we'll see some examples in Chapters 16 and 19. Second, in the absence of recombination, all of the variation in mtDNA is completely linked—mtDNA behaves as a single, haploid locus. This is both a blessing and a curse. It is a blessing because, as we shall see later, it is relatively straightforward to infer the history of a sample of mtDNA types when you don't have to worry about recombination. The only source of variation among mtDNA types then is mutation, and the number of mutations by which two mtDNA types differ directly reflects how long ago they last shared a common ancestor. That is, if I compare my mtDNA to your mtDNA and they differ by just one mutation, whereas the mtDNA of another person differs by 10 mutations, then you and I share a more recent common mtDNA ancestor with each other than we do with this other person (don't worry if this isn't clear, it will be explained in more detail in Chapter 11). We can use this principle to construct a **phylogeny** (basically, a genealogy) of the history of the mtDNA types in our sample, because a single phylogeny represents the history of the entire mtDNA genome. With recombination this can't be done, because different DNA segments then have different histories, so there isn't a single phylogeny for the entire DNA region.

The curse of mtDNA is that, as a single genetic locus, it is subject to the vagaries of chance and selection. Thus, the history of mtDNA may not reflect the history of a population or species, because of genetic drift or other chance fluctuations, or because there has been selection on mtDNA. To arrive at accurate inferences about population history, it is important to study many independent genetic loci (i.e., that are inherited independently), and we shall see some examples later where the picture of population history arising from genome-wide data differs from that arising from mtDNA analyses (e.g., whether or not we carry DNA from Neandertals). In fact, in this era of rapidly mushrooming genome-wide data, some have questioned whether there is still any value in analyzing mtDNA, but we will see some examples later where mtDNA analyses do provide useful insights. In particular, the comparison of mtDNA with Y chromosome variation is quite informative about sex-specific migrations and admixture, as shown in Chapters 16 and 19.

There are a variety of techniques for analyzing mtDNA variation. The earliest (pre-PCR era) studies analyzed RFLPs, either using the traditional Southern blot approach (described in Chapter 7) or using highly purified mtDNA and adding a radioactive label to the ends of the DNA fragments produced by restriction enzyme digestion in order to visualize them after electrophoresis (e.g., Brown 1980; Johnson et al. 1983). The analysis of mtDNA RFLP variation was greatly facilitated by the availability of the complete mtDNA genome sequence (all 16,569 nucleotides) in 1981 (Anderson et al. 1981), as it was then possible to infer in many instances the exact mutation responsible

for an observed RFLP. As with many other aspects of molecular genetics, the invention of PCR revolutionized studies of mtDNA variation, and the methodology of choice became direct sequencing of PCR products of the first hypervariable segment (HV1) of the mtDNA control region (Figure 9.2), an approximately 400 bp region that, as it's name suggests, contains a lot of mtDNA sequence variation. Such studies were also occasionally supplemented with sequencing another part of the control region, the second hypervariable segment (HV2), and/or genotyping of informative SNPs elsewhere in the mtDNA genome by methods such as PCR–RFLP or single-base extension assays (described back in Chapter 7).

In the past few years, sequencing of the entire mtDNA genome has become more common. This was first done by the traditional Sanger-sequencing approach (see Chapter 7) and involved a lot of time, effort, and money, as to do this meant amplifying the entire mtDNA genome from an individual in 24 or more overlapping fragments and then sequencing each PCR product (e.g., Rieder et al. 1998). For this reason, such studies usually screened samples first by HV1 sequencing and genotyping informative SNPs and then selected a subset of samples for complete mtDNA genome sequencing based on this prescreening. While this approach has the virtue of saving time and money, it does mean that the resulting sequences are not a random sample from the population, which thus limits or even precludes certain kinds of analyses (in particular, making demographic inferences, the subject of Chapter 12). Fortunately, the development of next-generation sequencing platforms has made it possible to quickly and efficiently determine complete mtDNA genome sequences from unbiased, random samples of individuals from populations (and at a fraction of the cost of Sanger sequencing). This, in turn, is making possible new kinds of demographic inferences from mtDNA genome sequences. While there are still only a few studies that have made use of this approach to date, given all the advantages it seems quite likely that next-generation sequencing of complete mtDNA genomes will soon become routine. For those of you interested in the details, an overview of one procedure (viz., the one we currently use) for carrying out next-generation sequencing of complete mtDNA genomes is provided in Figure 9.3.

The early RFLP-based studies of mtDNA variation revealed that the various mtDNA types could be grouped into **haplogroups**, based on the sharing of particular diagnostic mutations. The definition of a haplogroup is rather arbitrary, as at one extreme everyone could be defined as belonging to the same haplogroup (since all of our mtDNAs trace back to one common ancestor), while at the other extreme every unique sequence could be defined as a separate haplogroup. Moreover, the nomenclature for mtDNA haplogroups can be very confusing, as haplogroups were initially defined in a rather haphazard fashion as studies accumulated, rather than in any systematic fashion that reflects the actual mtDNA phylogeny. Thus, haplogroups A, B, C, and D are not closely related to one another as you might expect but rather happen to be the four haplogroups that predominate in the New World, as it was a study of mtDNA variation in the New World that first started naming haplogroups (Torroni et al. 1993). And the current way for naming haplogroups leaves a lot to be desired—one of the major haplogroups in Oceania, where I do a lot of work, has the highly inconvenient label of B4a1a1a (not to be confused with B4a1a1, also a major haplogroup in Oceania!). Still, the haplogroup system and nomenclature is so entrenched that there is little chance of any meaningful reform.

The phylogenetic relationships of the mtDNA haplogroups are shown in Figure 9.4, and a schematic view of their distribution in a worldwide sample of human populations (the CEPH–HGDP populations, discussed later in this chapter) is shown in Figure 9.5. Briefly, the first divergences in the mtDNA phylogeny involve haplogroups L0, L1, L2, and L3, all of which are found exclusively in Africa or in individuals with recent maternal African ancestry. Haplogroup L3 gave rise to two other major haplogroups called M and N, each of which in turn gave rise to several additional haplogroups (Figure 9.4). Essentially all individuals outside Africa (with the exception of those with recent maternal African ancestry) have mtDNA types that belong to one of the M or N haplogroups. What the phylogeny of mtDNA haplogroups has to tell us about origins of modern humans—in particular, the fact that the deepest divergences in the phylogeny are all in Africa, and mtDNA variation outside of Africa is a much-reduced subset of the mtDNA variation within Africa—will be explored in Chapter 16.

Y CHROMOSOMAL DNA

The paternal counterpart to mtDNA is provided by the poor, puny, pathetic little chromosome shown in Figure 9.6, and that of course is the Y chromosome, which is found only in males and transmitted from fathers to sons. Thus, analysis of Y chromosome variation gives insights into the paternal history of populations. There are only about a dozen or so genes on the Y chromosome, and not surprisingly most of them are involved in male fertility. Some regions of the Y chromosome do pair with and recombine with the X chromosome during meiosis, and these are called the **pseudoautosomal regions**. The rest of the Y chromosome does not undergo recombination and hence is sometimes

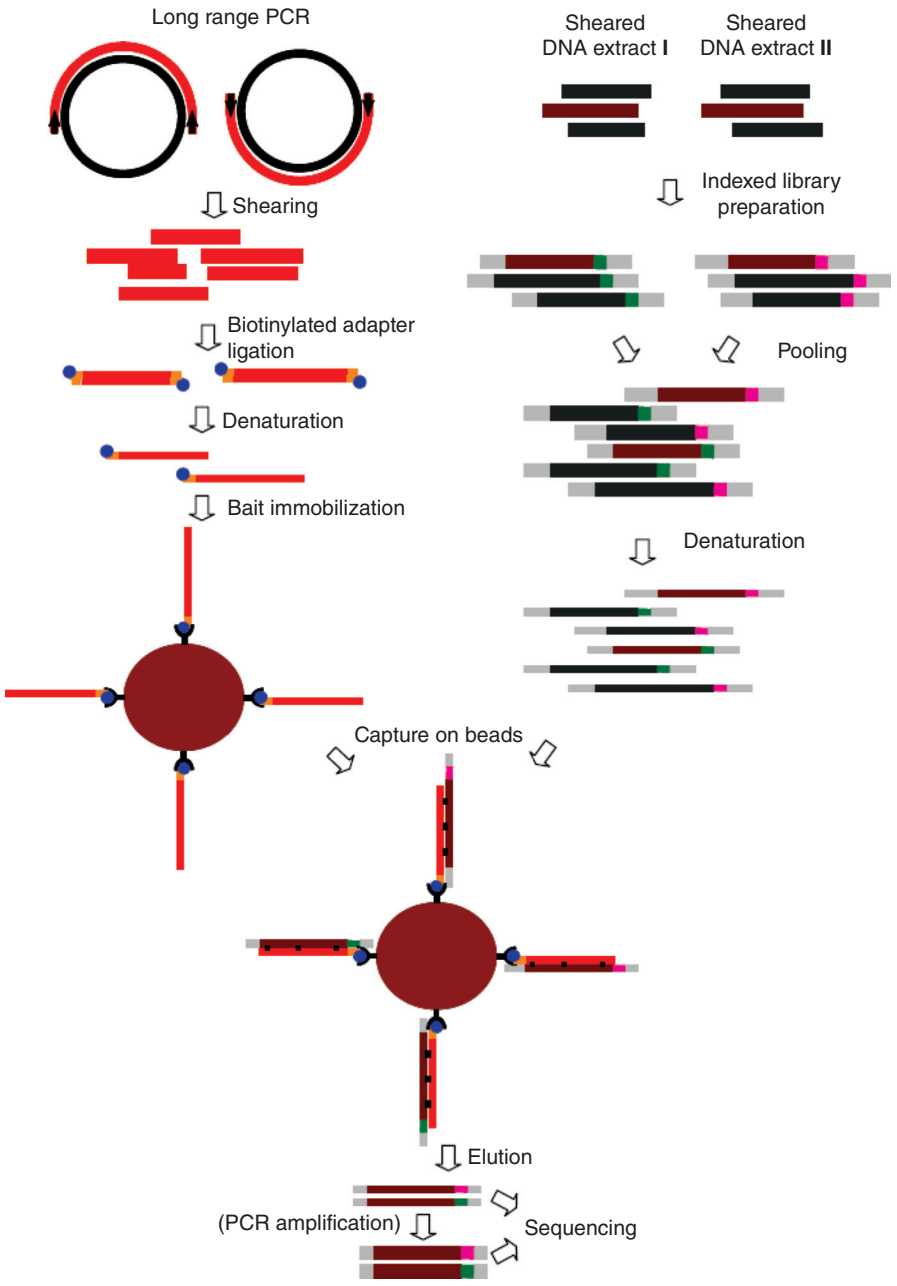


FIGURE 9.3

Overview of the capture method to enrich next-generation sequencing libraries for mtDNA sequences. Left, the "bait" is prepared by using long-range PCR to amplify the entire mtDNA genome in two overlapping segments from one sample. The PCR products are sheared and an adapter added to the fragments that can then be attached to magnetic beads—this is the bait. Right, samples for sequencing are processed by shearing genomic DNA and attaching indexes necessary for sequencing and to uniquely label each library. Up to 96 samples are then pooled, denatured, and hybridized to the magnetic beads, which capture mtDNA sequences (based on the principle of DNA complementarity). The beads are washed to remove sequences that do not hybridize to the bait, then the captured sequences are eluted from the beads, amplified by PCR to produce enough DNA for sequencing, and then sequenced on a next-generation sequencing platform. PCR, polymerase chain reaction. Reprinted with permission from Maricic, T., et al., "Multiplexed DNA sequence capture of mitochondrial genomes using PCR products," *PLoS ONE* 5:e14004, 2010.

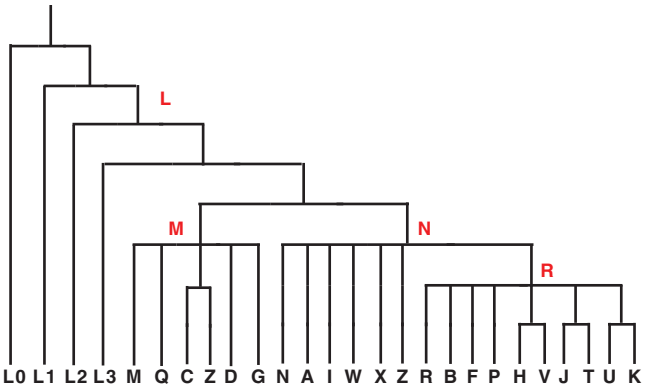


FIGURE 9.4

Phylogenetic tree illustrating the relationships of the major mtDNA haplogroups. Macrohaplogroups L, M, N, and R are indicated. Note that branch lengths are not proportional to mutational differences.

referred to as the nonrecombining Y chromosome, or NRY for short. The NRY consists of a few unique sequence regions, which can be readily analyzed using a variety of methods, as well as many regions that consist of various copies of long repeats. There is considerable variation among Y chromosomes with respect to these repeats and their number and orientation (i.e., direct or inverted), which has hampered investigation of variation in these repeat regions (I will refrain from any sexist comments as to why it is that only males have a chromosome with such a messy structure!).

For many years, it was thought that the NRY harbored little or even no sequence variation in humans; indeed, as recently as 1995 it was even possible to publish a study on the lack of NRY variation in humans in the prestigious journal *Science* (Dorit et al. 1995). This picture changed dramatically just a few years later with the development of more sensitive techniques for discovering SNPs, and led largely by the efforts of the "two Peters," namely, Peter Oeffner and Peter Underhill (working with the legendary Luca Cavalli-Sforza),

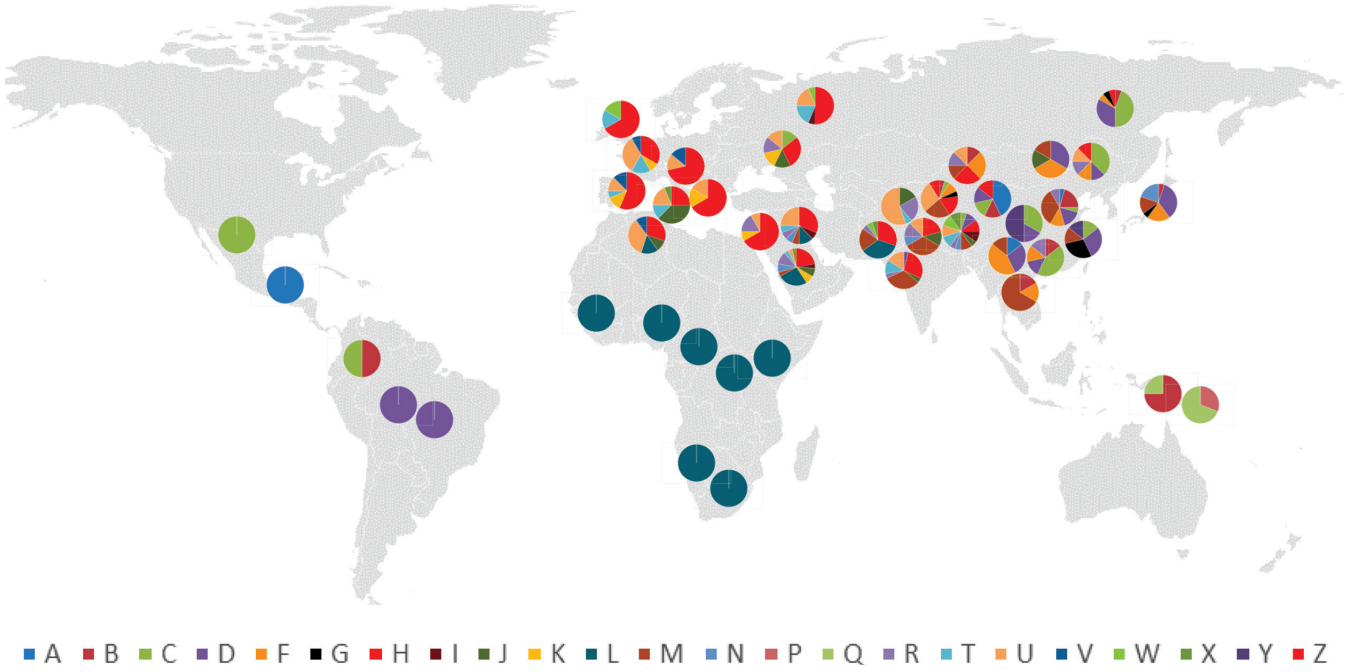


FIGURE 9.5

Distribution and frequency of the major mtDNA haplogroups in the CEPH–HGDP populations. To enhance visual clarity, some populations from China and Pakistan are not included, and the positions of some populations have been shifted slightly. This figure should be viewed with a healthy dose of skepticism, as the CEPH–HGDP does have significant gaps in the sampling of populations, and sample sizes for most populations are quite small. Moreover, the designation of "major" haplogroup has a strong Eurocentric bias, as it looks as if Eurasia has many major mtDNA haplogroups while sub-Saharan Africa has just one. In fact, if haplogroups actually reflected the time-depth of the corresponding mtDNA lineages, then Africa would have many major haplogroups, and all non-Africans would fall into just two haplogroups, M and N (see Figure 16.3). Nonetheless, the figure serves to give an overall impression of the geographic distribution of mtDNA haplogroups. The data used to create this figure are from Lippold, S., et al., "Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences," *Investigative Genetics* 5:13, 2014.

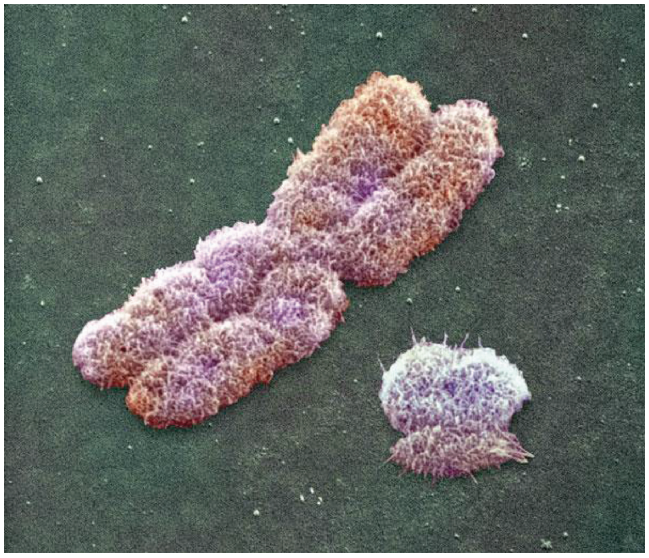


FIGURE 9.6

Electron micrograph of a human Y chromosome (right), in comparison to a human X chromosome (left). Reprinted with permission from Willard, H., "Tales of the Y chromosome," *Nature* 423:810, 2003.

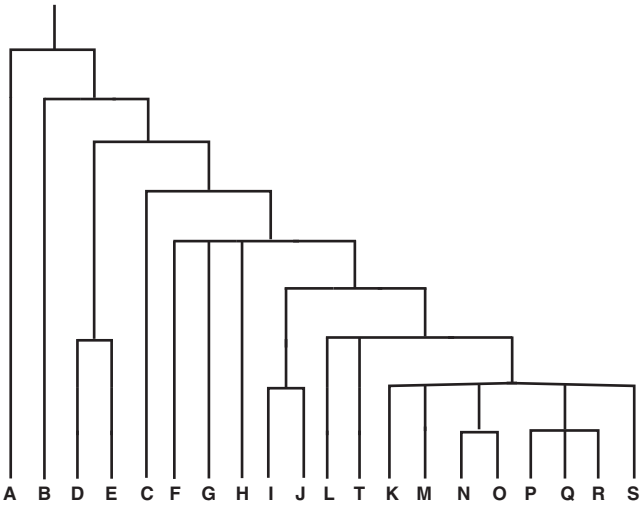


FIGURE 9.7

Phylogenetic tree illustrating the relationships of the major Y chromosome haplogroups. As with the tree of mtDNA haplogroups (Figure 9.4), the branch lengths are not proportional to mutational differences.

hundreds of SNP markers soon became available (Underhill et al. 2000). With numerous STR (short tandem repeat) markers also becoming available, NRY analyses quickly came into their own and are now an established tool of molecular anthropology studies. Moreover, next-generation sequencing platforms are enabling the determination of partial Y chromosome sequences via capture-array–based methods (e.g., Lippold et al. 2014), which in turn allows for richer analyses, as the Y chromosome sequences are free from ascertainment bias (unlike the SNP-based genotype data). Like mtDNA, the NRY is a single, haploid genetic locus, with the same strengths (e.g., relatively easy to infer the phylogeny of NRY types) and weaknesses (e.g., highly subject to the vagaries of genetic drift). As we shall see in Chapters 16 and 19, analyzing both mtDNA and NRY variation to get at both the maternal and paternal history of populations has provided many important insights.

As with mtDNA, different NRY variants are classified into haplogroups on the basis of diagnostic mutations. However, the people who decided on the nomenclature for NRY haplogroups did learn a lesson as to what not to do from the mtDNA haplogroup nomenclature, and as a result the NRY haplogroup nomenclature is organized phylogenetically (Figure 9.7). That is, the first split in the NRY phylogeny involves haplogroup A, followed by haplogroup B, and so forth. The distribution of NRY haplogroups in the same populations depicted in Figure 9.5 is shown in Figure 9.8; the astute reader may notice that there appears to be somewhat more geographic specificity (i.e., less sharing among populations and hence bigger differences) for NRY haplogroups than for mtDNA haplogroups. A possible explanation for this difference will be forthcoming in Chapter 19. The well-informed reader may already know about a complication to the Y chromosome phylogeny known as haplogroup A00; this will be discussed in a later chapter.

AUTOSOMAL DNA

Compared to mtDNA and the NRY, the two chief features of the autosomes are the vastly greater amount of genetic information that potentially can be exploited, and the fact that recombination occurs during meiosis, thereby mixing up the maternal and paternal chromosomes each generation. Depending on how much of the mtDNA genome is sequenced, one can expect to find tens to hundreds of polymorphic sites in a typical study, and most NRY studies will analyze on the order of 10–50 SNPs and/or a dozen or so STR loci (or a few thousand SNPs in the new sequence-based studies); by contrast, as we saw in the chapter on genetic markers, SNP chips enable genotyping of hundreds of thousands to a few million or so SNPs, while full genome sequences provide millions of polymorphic positions. As we shall see in subsequent chapters, having vastly more data enables novel types of analyses that cannot be carried out with mtDNA or NRY variation alone. And, having lots and lots of independent loci means that a more accurate picture of the overall genetic

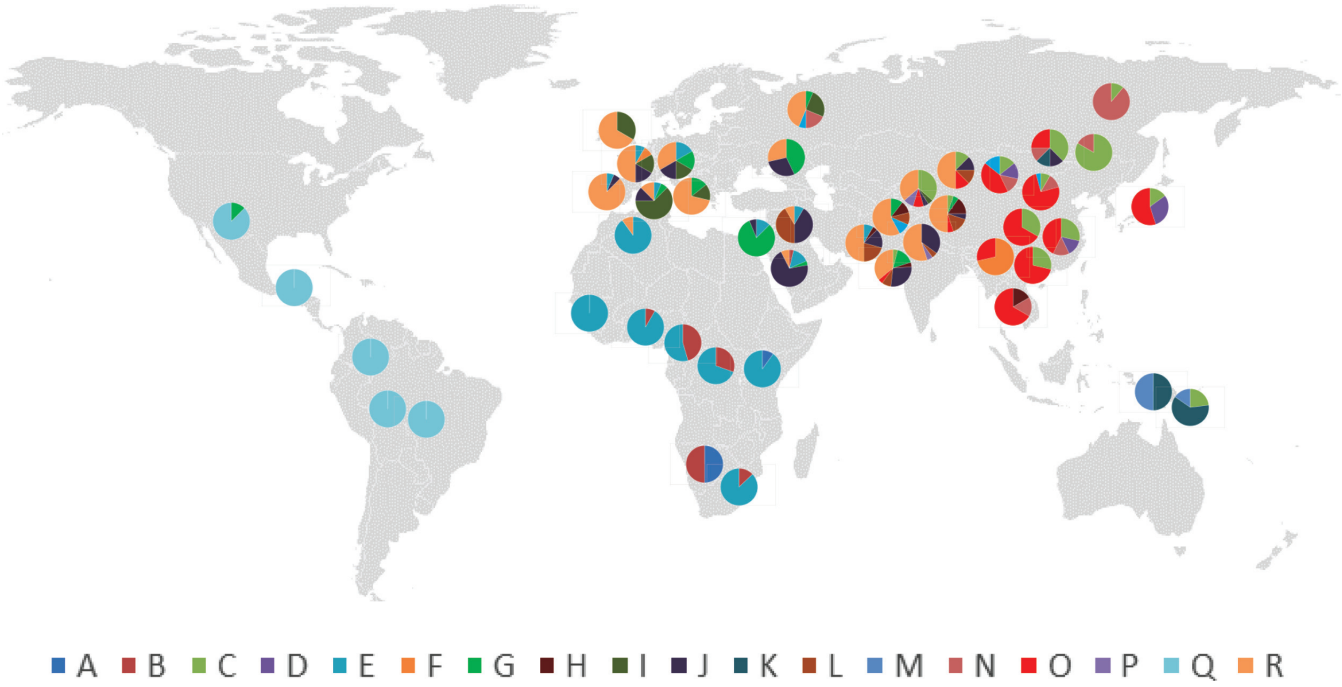


FIGURE 9.8

Distribution and frequency of the major Y chromosome haplogroups in the CEPH–HGDP populations. All of the caveats mentioned in the legend to Figure 9.5 (the map of mtDNA haplogroups in the CEPH–HGDP populations also hold here or even more so, e.g., sample sizes are even smaller since the data are limited to the males). Nevertheless, the astute reader who compares the mtDNA and Y chromosome haplogroup distributions may get the impression that Y chromosome haplogroups seem to differ more among populations than do mtDNA haplogroups; this observation will be discussed in more detail in Chapter 19. The data used to create this figure are from Lippold, S., et al., "Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences," *Investigative Genetics* 5:13, 2014.

history of a population can be obtained than from just mtDNA and/or NRY analyses. This is because the mtDNA and the NRY each behave as a single genetic locus and hence like any single locus may have been influenced by genetic drift (as discussed in Chapter 5). Autosomal loci are of course also subject to genetic drift, but by averaging across many loci the effects of genetic drift on the results can be minimized—it's like the difference between flipping one coin 10 times to estimate the probability of a heads versus flipping several thousand coins 10 times each and taking the average of the outcomes.

Although it used to be thought that recombination was a problem because it complicated the analysis of autosomal DNA data (compared to mtDNA or NRY data), in the past few years it has been recognized that in fact recombination (or, more accurately, associations between linked loci) can provide additional insights into population history. Instead of just analyzing the data as a large number of independent SNPs, by investigating the associations between genotypes at closely linked SNPs (i.e., SNPs that are physically located near each other on the same chromosome) you can actually get more information—in other words, the whole is greater than the sum of the parts.

Nonrandom association between genotypes at linked SNPs is known by the cumbersome name of **linkage disequilibrium**, or LD for short. Don't be confused by the difference between linkage and LD: as we saw in Chapter 1, linkage refers to pairs of loci that are close enough together on the same chromosome that they violate Mendel's Law of Independent Assortment—alleles at such loci are inherited together more often than expected by chance. Linkage relationships are, therefore, based on observations from families, while LD, on the contrary, is based on nonrandom associations between alleles in populations. If I have one locus with alleles A_1 and A_2, each at a frequency of 0.5, and another locus with alleles B_1 and B_2, also at frequencies of 0.5 each, then in the absence of LD, the frequency of A_1B_1 chromosomes is just $(0.5) \times (0.5) = 0.25$. More generally, if $f(A_1)$ is the frequency of the A_1 allele and $f(B_1)$ is the frequency of the B_1 allele, then with no LD, the expected frequency of A_1B_1 chromosomes is simply $f(A_1) \times f(B_1)$. So, if the observed frequency of A_1B_1 chromosomes departs

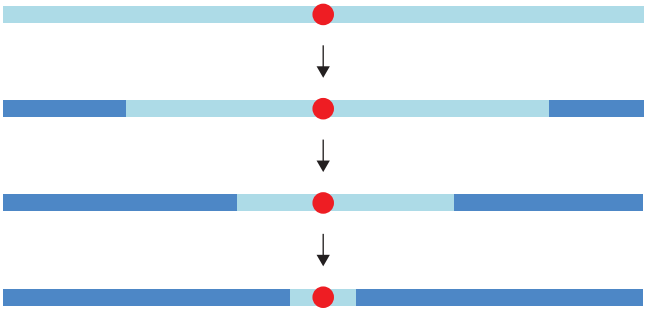


FIGURE 9.9

Over time, the association between a newly arisen mutation (red circle) and the original haplotype (light blue) decreases due to recombination introducing new segments (dark blue).

significantly from this expected frequency, then by definition there is significant LD between the A and B loci.

How can LD arise? The basic idea behind LD is illustrated in Figure 9.9. As the figure indicates, consider what happens to a new mutation. By definition, that mutation exists as just a single copy in the population, on a single chromosome, in complete association with all of the alleles on that particular chromosome. The **haplotype** (set of associated variant alleles) for a new mutation, therefore, consists of the entire chromosome, and the new mutation is in complete LD with all other loci on the chromosome. Suppose this new mutation is one of the lucky few that is not lost immediately by drift (which, you will recall from Chapter 5, is the usual fate of most new mutations) but instead starts to increase in frequency in the population. As shown in Figure 9.9 as the new mutation increases in frequency, the length of the haplotype associated with the new mutation will get shorter and shorter over time, because recombination (as well as new mutations at other loci on the chromosome) will introduce different alleles onto the chromosomes carrying this new mutation. Eventually, if recombination goes on for a long enough period of time, this new mutation will no longer be associated with specific alleles at the other loci on the chromosome, and hence there will be no LD.

So, a new mutation starts out in complete LD and then the LD decays over time due to recombination and mutations at other loci. And since the amount of recombination and new mutation is (partly) dependent on the demographic history of the population (changes in population size, migration from other populations, etc.), we can use information on LD to make inferences about population history. Moreover, the rate of change of LD also differs between neutral alleles and alleles that have been subject to positive selection; as we shall see in Chapter 18, this property has been exploited by novel methods to detect positive selection from genome-wide data. And, as we shall see in Chapter 12, when a population experiences admixture with another population with a different ancestry that can be distinguished in genome-wide data, then the current size and number of haplotypes (sometimes called **admixture blocks**) in our population of interest that come from this other population can be used to infer how much admixture occurred and when it occurred. Linkage disequilibrium (and its decay) is thus very informative about many aspects of population history that we would like to know about, and new methods that exploit LD are being developed all the time. And again, because there is no recombination in mtDNA or the NRY, there is no possibility of applying methods based on LD decay to mtDNA/NRY data. However, it should also be kept in mind that LD-based methods for making inferences about population history do require dense genome-wide data on the order of tens of thousands to hundreds of thousands of SNPs—they cannot be applied to just any autosomal DNA data set.

X CHROMOSOME DNA

When it comes to studies of population history, the X chromosome has largely taken a back seat to mtDNA, the NRY, and the autosomes. In fact, many studies that utilized commercially available microarrays ("SNP chips") for genotyping obtained X chromosome data (because X chromosome SNPs are on the chips) but simply ignored the X chromosome data in the analyses. This is because if you have a mixture of males and females in your study, then for the X chromosome you have a mixture of haploid (from the males) and diploid (from the females) genotypes, and trying to include both sorts of data in a single analysis is quite complicated.

And yet, there are a number of advantages to analyzing X chromosome data. Many analyses require knowing the **phase** of the data—that is, which alleles at heterozygous loci go together on each of the two copies of a chromosome. The concept of phase was introduced way back in Chapter 1, but to recapitulate, suppose we have two SNPs on the same chromosome, and an individual is heterozygous A_1A_2 at the first SNP and B_1B_2 at the second SNP. Then the two possibilities are A_1B_1 on one chromosome and A_2B_2 on the other, or A_1B_2 on one chromosome and A_2B_1 on the other. How can you tell which of these is the actual phase? If you have family data, you can (sometimes) figure out the phase from the genotypes of the parents. In this example, if the parents of our individual had genotypes A_1A_1, B_2B_2 and A_2A_2, B_1B_1, then we would know that the phase was A_1B_2 on one chromosome and A_2B_1 on

the other (but note that if both parents had genotypes A_1A_2, B_1B_2, then we could not figure out the phase).

However, in molecular anthropology studies we seldom have family data (especially if, as mentioned in Chapter 8, we try to avoid relatives when sampling individuals). We, therefore, usually have to make use of algorithms that examine all of the genotypes in the individuals studied and then try to figure out the most likely combination of phased chromosomes that explain the data. In the above example, if A_1 is much more frequent than A_2, and B_1 is much more frequent than B_2, then we might observe that most individuals in our sample have genotypes A_1A_1, B_1B_1; A_1A_2, B_1B_1; or A_1A_1, B_1B_2. If so, then it is reasonable to infer that we have three haplotypes in our sample: A_1B_1, A_2B_1, and A_1B_2, and, therefore, the most likely phase for our A_1A_2, B_1B_2 individual is A_1B_2 and A_2B_1. Looks reasonable—but in practice, phasing genome-wide data is computationally very intensive (i.e., it can take weeks or even months of computer time for hundreds of thousands of SNPs and lots of individuals) and error-prone in that one mistake in the phasing can influence many loci on a chromosome. While there are some new experimental approaches on the horizon that will provide the phase directly—for example, sequencing approaches that can generate long sequences from a single molecule (e.g., Carneiro et al. 2012)—at the moment, these computational approaches to phasing are the best that can be done.

But for the X chromosome, there is another solution to the phasing problem: as males have a single X chromosome, the phase is known without error for X chromosome data from males, which is a tremendous advantage. Moreover, comparison of patterns of variation between the X chromosome and the autosomes can, like comparisons between maternally inherited mtDNA and paternally inherited NRY variation, provide insights into sex-specific processes that influence human genetic variation. This is because the average X chromosome spends twice as much time in females as in males, so X chromosome variation reflects the maternal history to a greater extent than the paternal history of populations. Still, analysis of X chromosome variation is not without complications—for example, it appears that selection may have influenced the X chromosome more than the autosomes, especially for selectively advantageous recessive mutations. As we saw in Chapter 5, a new, advantageous recessive mutation on the autosomes is initially at the mercy of drift, as it must reach a high enough frequency for homozygotes to occur for the advantageous phenotype to be exhibited and hence for selection to occur. But any new, advantageous recessive mutation on the X chromosome will immediately exhibit the associated phenotype in any males with the mutation (because of the hemizygous nature of the X chromosome) and hence selection will be all that more effective. Overall, then, while X chromosome studies are on the rise, the X chromosome has lagged behind mtDNA, NRY, and autosomal variation when it comes to studying population history.

PUBLIC DATABASES

So far, our discussion of sampling has revolved around issues related to sampling individuals/populations and the properties of various types of DNA. There is a third piece to the puzzle and that is the methods that one uses to analyze and make inferences from the data that are generated—you may have the ideal samples and DNA data for what you want to know, but these won't matter if you don't also carry out the appropriate analyses of the data (especially if the way you obtained your samples and/or the properties of the genotypes you collected violate some vital assumption of the analyses you want to do!). We will turn to this important topic in the next few chapters.

However, before leaving the topic of sampling issues, there is one final aspect to discuss and that concerns public databases of genetic data. As an alternative to going out and collecting samples and carrying out the laboratory analyses, it is also possible to analyze the DNA data available in public repositories. This approach has the obvious advantage of being much easier, faster, and cheaper than mounting an expedition to collect samples and produce genetic data in the laboratory. It also has obvious limitations in that you can only analyze whatever is available—if the public databases don't include samples and/or DNA regions that you need to address the question(s) that interest you, too bad. And you also have to beware of inappropriate use of the data—recall the example at the beginning of Chapter 8 concerning the three HapMap populations from Africa, Asia, and Europe, and how some studies assumed that the results from these three populations could be extrapolated to the entire human species. Most of the public repositories were established to facilitate health and disease-related research, which needs to be kept in mind when using their data for molecular anthropology studies. Still, the public repositories have provided an extraordinarily valuable source of information and insights; for example, many new methods for making inferences about population history or selection were developed and tested on publicly available data, which helps serve as a benchmark for comparing the performance of different methods.

There are many public repositories that can readily be found by an Internet search, but a few do merit pointing out here, including:

Genbank (http://www.ncbi.nlm.nih.gov/genbank): this is the granddaddy of them all, run by the US National Institutes of Health and containing essentially all publicly available DNA sequence data. Established in 1982, there are currently more than 150 million sequences in Genbank—if it's been sequenced and published, you'll probably find it there. The sequences are annotated, meaning various descriptive features are listed, including species and population origin, making it straightforward to find out what is available for your favorite population or segment of DNA.

HapMap (http://hapmap.ncbi.nlm.nih.gov): the International Haplotype Map Project (or HapMap for short) was established in 2002 with the goal of providing a catalog of common variation in human populations that would aid in finding genes associated with disease. Since, as we saw in the above section on autosomal DNA, human genetic variation is not distributed at random across chromosomes but instead is organized into haplotypes, if you know the genotype of a SNP that is associated with a particular haplotype (called a **tag SNP**), then you know the haplotype and the corresponding genotype of all the other SNPs in that haplotype without having to do any additional genotyping. The idea behind HapMap, therefore, is that by identifying common haplotypes and corresponding tag SNPs in a handful of populations from around the world, this information can be more widely used to identify haplotypes associated with particular diseases in other populations. HapMap started with four populations (Yoruba from Nigeria, European–Americans from Utah, Han Chinese from Beijing, and Japanese from Tokyo, with the latter two often combined into one East Asian sample) genotyped for about 1 million SNPs. The most current release consists of 1.5–4 million SNP genotypes for 11 populations, with more on the way. In addition to making the data available for download, the HapMap Web site also includes some useful tools for finding SNPs in a particular DNA region of interest and for identifying tag SNPs.

1000Genomes (http://www.1000genomes.org): launched in 2008, the 1000Genomes project is an international project with the seemingly ambitious goal of providing 1000 full genome sequences from a sample of humans from around the world, sufficient to identify every genetic variant at a frequency of 1% or more in the populations studied. Thanks to advances in next-generation sequencing technology, this goal has been revised to (at least) 2500 full genome sequences from (at least) 25 populations. Utilizing the HapMap populations as the primary resource, currently a mix of data are available for more than 1000 individuals, consisting variously of high-quality (~30X, meaning that each nucleotide in the genome is sequenced on average 30 times) genome sequences, low-quality (~4X) genome sequences, and exome sequences (i.e., sequences of all of the exons, obtained by capture hybridization using arrays that contain probes to all of the exons). In addition to making all of the sequence data publicly available, the 1000Genomes Web site provides a number of useful tools for exploring and visualizing the sequence data.

HGDP: Beginning in the early 1990s, the late Allan Wilson, Luca Cavalli-Sforza, and other scientists called for a Human Genome Diversity Project (analogous to the then ongoing Human Genome Project to sequence the complete human genome), an ambitious attempt to systematically sample and study genetic diversity in human populations from around the world. Several planning meetings were held, but the project fell afoul of political and ethical concerns, with some organizations that represent the concerns of indigenous peoples accusing the project of propagating racist views and Eurocentric exploitation of indigenous people. The project was abandoned as such, but what came out of this effort was a decision by the Centre d'Étude du Polymorphisme Humain (or Human Polymorphism Study Center, also known as CEPH) in Paris to establish a collection of immortalized cell lines from worldwide human populations and to make available DNA from this collection to investigators for a nominal fee. Usually, when living cells are grown in culture, after about 50 or so generations (cell divisions) they stop growing and die; by contrast, immortalized cell lines can be grown forever and are commonly obtained by treating white blood cells **(lymphoblasts**) with a virus. Such cell lines are an inexhaustible source of DNA, as the cells can be frozen and then revived whenever needed. Announced in 2002, the CEPH Human Genome Diversity Panel (HGDP) consists of more than 1000 samples from 52 worldwide populations (Figure 9.10)—it is thus unique among the public repositories discussed here in that it is the only one designed with molecular anthropology and population history studies in mind. Still, the HGDP is hardly a representative sampling of the world's populations, as several important regions (such as Australia) are not represented at all, whereas other regions are overrepresented, like Pakistan with eight groups in the HGDP. These vagaries of sampling reflect what was available in the way of cell lines that other investigators were willing to donate and for which the necessary ethical permission had been obtained to have DNA samples distributed to other scientists. Given the difficulties associated with sampling to produce immortalized

FIGURE 9.10

Map showing the location of the populations included in the CEPH–HGDP panel of DNA samples. From Young, J.H., et al., "Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion," *PLoS Genetics* 1:e82, 2005.

cell lines—you need live blood cells, which means having freshly drawn blood that must be kept cold but not frozen and has to be back in the laboratory for processing within about 48 hours—it is a very impressive collection of DNA samples that has proven extraordinarily useful in dozens of studies, as we shall see in later chapters. And a major reason for the widespread use of HGDP samples is not just their ready availability but also the fact that investigators who make use of HGDP samples are required to deposit their results in a central database, available from the CEPH–HGDP Web site. Currently, the database consists of more than 1 million genetic markers (SNPs, indels, CNVs, microsatellites, etc.) with much more to come—already full genome sequences have been produced for a few HGDP samples, and one can anticipate, as sequencing costs drop, that eventually full genome sequences will be available for all of the samples.

One final point about public data sets: an unanticipated but extremely important contribution is that they have helped "democratize" the way molecular anthropology is done. Thanks to the widespread availability of genetic data, you don't need the wherewithal or funding to mount a sampling expedition and run a laboratory—in fact, you don't even need a Ph.D. degree or an academic position—to do research in molecular anthropology these days. And not only are all sorts of data readily available but many software packages for carrying out various types of analyses are also available for free—or, for the truly ambitious, you can write your own analysis programs. All you need is an Internet connection and a good idea, and there are several Web sites and blogs where analyses are carried out and discussed by enthusiastic "amateurs," all on the basis of publicly available data and methods.

CHAPTER 10

ANALYSIS OF GENETIC DATA FROM POPULATIONS

So now you've sampled your groups and done the laboratory work—or, you've downloaded some genotype or sequence data from one of the public repositories—and you're eager to see what you can learn about the genetic history of your populations of interest. How, exactly, do you do that? Analyzing and making inferences from genetic data is the subject of the next three chapters. In this chapter, we will discuss analyses of genetic variation within populations and genetic differences among populations; the emphasis is on the population as the unit of study. In the following two chapters, we will discuss analyses where the individual (based on either genotype or DNA sequence data) is the unit of study and then we will discuss various ways to infer demographic history from genetic data. The emphasis in all three chapters is on population history, that is, questions such as: what are the genetic relationships of the population(s)?; where did the ancestors of this population come from?; when did this population diverge from its nearest relative?; did any migration occur in the past (and if so, when and how much)?; what is the history of population size changes?; and so forth. Analyses that concern other aspects of molecular anthropology, such as detecting natural selection, will be covered later.

GENETIC DIVERSITY WITHIN POPULATIONS

Let's suppose we have determined mtDNA sequences from 10 individuals from each of two groups, and in the first group everyone has the same mtDNA sequence, while in the second group everyone has a different mtDNA sequence. Obviously, the second group has lots more genetic diversity than the first one, but how can we make this more quantitative? One way that we have already seen is to ask what is the probability that two alleles (in this case, mtDNA sequences) drawn at random from a group are different. This, you will recall from Chapter 4, is one way we can think about heterozygosity in the context of Hardy–Weinberg: if we have alleles $A_1, A_2, \ldots, A_n$ each with frequencies $x_1, x_2, \ldots, x_n$ (so the sum $x_1 + x_2 + \ldots + x_n = 1$), then $\sum x_i^2$ is the probability that two alleles drawn at random are the same, so $1 - \sum x_i^2$ is the probability that two alleles drawn at random are different. Even though, strictly speaking, there are no heterozygotes for mtDNA (because mtDNA is haploid), this is still a useful measure of comparative genetic variability that can be applied to haploid as well as diploid data. So in our first group where everyone has the same sequence, we have one allele with a frequency of 1, and the heterozygosity is 0. And in the second group, where everyone has a different sequence, we have 10 alleles, each with a frequency of 0.1, and so the heterozygosity is 0.9. And if we imagine a third group of 10 individuals, with a total of two different sequences each present in five individuals, then the heterozygosity is 0.5, which hopefully makes sense—clearly, this group has more genetic variation than the group where everyone has the same sequence but less genetic variation than the group where everyone has a different sequence, so we get an intermediate value for the heterozygosity.

So far, so good, but let's consider another sample of 100 individuals, each with a different sequence. Then we have 100 alleles, each with a frequency of 0.01, and so if you do the math you find that the heterozygosity is 0.99 for this group. Now our measure of genetic variation doesn't seem so satisfactory: two groups have the same overall variability (i.e., everyone has a different sequence), but we get different heterozygosity values simply because the sample sizes are different. What can we do about this? If you stare at these numbers long enough, you realize that when $n = 10$, the heterozygosity is 9/10, and when $n = 100$, the

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

heterozygosity is 99/100. So, in both cases, the heterozygosity is $(n-1)/n$, and it turns out that this will always be the case whenever everyone in your sample has a different allele—if you don't believe me, try a few other values for n (or, if your algebra is up to the challenge, replace x_i by $1/n$ in the equation for heterozygosity and see what you get). Then, if we multiply the heterozygosity value by $n/(n-1)$ for the group with 10 individuals, we get $0.9 \times (10/9) = 1$, and for the group with 100 individuals, we get $0.99 \times (100/99) = 1$. Much better—when everyone in our sample has the same allele, we get a value of 0, and when everyone has a different allele, which intuitively seems like the maximum amount of genetic variation you can have, then we get a value of 1, regardless of the sample size. This equation:

$$\left(1 - \sum x_i^2\right) n/(n-1)$$

is known as the **genetic diversity** value and is often designated by H. Genetic diversity is preferred over the term heterozygosity because even though H is clearly related to heterozygosity, the equation holds for haploid systems such as mtDNA and the NRY although there are no heterozygotes as such. Note that n is the number of alleles or chromosomes in the sample; for mtDNA and the NRY, this is the same as the number of individuals sampled, but for autosomal loci, the number of alleles is twice the sample size. The term $n/(n-1)$ is also known as the correction for sample size and comes up frequently when estimating statistics like genetic diversity from real data; note also that the bigger the sample size, the closer the correction value is to 1 (and hence the less important it becomes).

While H is certainly a useful and widely used measure of genetic variation, it does not tell you how different the alleles are at the molecular level. That is, if in one group there are 10 mtDNA sequences that are all different and differ from each other by two mutations on average, and in another group there are 10 mtDNA sequences that are also all different but differ from each other by 10 mutations on average, the H values for both samples will still be 1. And yet, intuitively the population with 10 differences on average between mtDNA sequences would seem to have more genetic variation than the population with two differences, so we would like to take this into consideration. We could simply count the number of polymorphic sites, but this also has the drawback of being dependent on sample size—in general, the bigger the sample size, the more polymorphic sites you expect to detect. One measure of genetic variation that neatly gets around this is the mean number of pairwise differences, or MPD for short. As the name suggests, this is easy enough to calculate: take every pair of sequences in your sample, count how many differences there are,

FIGURE 10.1

Example of how to calculate the MPD from a set of sequences. Starting with the five sequences (**A**), count the number of differences between each pair of sequences (**B**), then sum these and divide by the number of pairs of sequences to get the MPD (**C**). MPD, mean number of pairwise differences.

then sum the number of differences across all pairs of sequences and divide by the total number of pairs (which is $n(n-1)/2$) to get the mean number of pairwise differences (see Figure 10.1 for an example). This measure applies only to situations in which one can actually (and easily) count the number of mutational differences between individual alleles, so it is applied to DNA sequences, NRY haplotypes based on STR (short tandem repeat) loci, and so forth, but not to allele frequency data.

One problem with MPD is that the number of pairwise differences observed in sequence data will also depend on how many bases are sequenced. If we are sequencing the same DNA segment (e.g., mtDNA or the same portion thereof), then there is no problem, but if we want to compare diversity estimates for different DNA segments, then this becomes an issue—the longer the DNA segment, the more polymorphic sites you expect, and hence the more pairwise differences between individuals. To account for differences in sequence length, one can simply divide MPD by the length of the corresponding sequence, and this measure is known as the **nucleotide diversity** per site and is usually designated as π. Nucleotide diversity is a highly useful measure of genetic diversity, not only because it can be directly compared between studies even when sample sizes and/or DNA segment lengths differ but also because it turns out to be an estimate of $4N\mu$ (also known as Θ), where N is the effective population size and μ is the neutral mutation rate. Why is this important? Recall that we discussed this back in Chapter 5 in the section on the equilibrium between neutral mutations and genetic drift and how Θ is related to the expected amount of genetic variation in a population. And, as we shall see in Chapter 17, there are different ways to estimate Θ that are the basis

TABLE 10.1 ■ Diversity values for different markers and populations[a]

	Classical markers[b]	RFLPs[c]	STRS[c]	mtDNA[c]	Nuclear DNA[d]
Africans	0.208	0.322	**0.769**	**0.031**	**0.00093**
Asians	0.212	0.377	0.681	0.011	0.00081
Europeans	**0.262**	**0.432**	0.724	0.010	0.00076

[a] H values are given for classical markers, RFLPs, and STRs, while π values are given for mtDNA and nuclear DNA sequences. The values in **bold** indicate the population with the highest diversity value for each type of marker.
[b] 23 loci (Nei and Roychoudhury 1982).
[c] 30 RFLP loci, 30 STR loci, 270 bp HV2 mtDNA sequence (Jorde et al. 1995).
[d] 21 kb (Yu et al. 2001).

for a widely used test to see whether genetic data fit the predictions of neutrality. Note that π is an estimate of $4N\mu$ for autosomal data, but for mtDNA, it is an estimate of $2N_f\mu$, where N_f is the effective population size for females (the reduction by a factor of 2 comes about because mtDNA is haploid), and similarly for the NRY, it is an estimate of $2N_m\mu$, where N_m is—you guessed it—the effective population size for males.

Let's have a look at some diversity values for different genetic markers in some human populations. Table 10.1 shows H values for classical markers, RFLP loci, and STR loci and π values for mtDNA and nuclear DNA sequences, for Africans, Asians, and Europeans. Note first that there is great variation among the estimates for different types of markers—the H values for classical markers are lower than those for RFLPs, which in turn are lower than those for STR loci, while π values are higher for mtDNA than for nuclear DNA sequences. These results should not come as a surprise to you, if you remember the discussion of the properties of various types of genetic markers (in Chapter 7) and genomic regions (in Chapter 9). In particular, recall that classical markers show only variation for amino acid substitutions that result in differences either in the antigenic properties of immunological markers such as blood groups or in the electrophoretic mobility of proteins. By contrast, RFLPs are based on polymorphisms at the DNA level that influence restriction enzyme recognition sequences, and, therefore, it is not surprising that RFLP markers reveal more variation than classical markers. Similarly, STR markers, which vary in the number of copies of tandem repeats, have a much higher mutation rate than the nucleotide substitutions that underlie either classical markers or RFLPs, so it is not surprising that STR markers have higher H values. Moreover, since mtDNA has a higher mutation rate on average than nuclear DNA, it is not surprising that mtDNA has higher π values than nuclear DNA.

What should come as a surprise, though, is the observation that different populations have the highest diversity values for different markers. Namely, Europeans have the highest diversity values for classical markers and RFLPs, while Africans have the highest diversity values for STRs and mtDNA and nuclear DNA sequences. What's going on? The first thing you should be suspicious about in any sort of comparison like that shown in Table 10.1 is the populations—what, exactly, is meant by such vague labels as "Africans," "Europeans," and "Asians," and were the same individuals studied for all of these different kinds of markers? As it turns out, different groups of individuals were studied for the different markers (except that the RFLPs, STRs, and mtDNA sequences were obtained from the same individuals). So, in principle, sampling of different populations could account for at least some of this variation in terms of which population has the most diversity for a particular kind of marker. However, I will ask you to take my word for it that the variation among diversity values that has been observed for different populations from within these continental regions (i.e., Africa, Asia, and Europe) is highly unlikely to account for the results in the table—and if you are still skeptical on this point, do note that the exact same individuals were analyzed for RFLP, STR, and mtDNA sequence variation, yet Europeans have the most RFLP variation while Africans have the most STR and mtDNA sequence variation. So, differences in population sampling cannot account for the differences in diversity.

It turns out that the answer has to do with how the particular markers were chosen for analysis. By their very nature, classical markers or RFLPs have already been determined to be polymorphic in at least one population. If a particular red cell antigen does not exhibit any variation, then there won't be any variation in terms of who has antibodies against this antigen and who doesn't, and so the existence of this antigen will go undetected. And while in principle it would be possible to survey electrophoretic variation in proteins, as well as restriction enzyme variation in particular DNA segments, without knowing whether they are polymorphic or not, in practice it costs valuable time and money to carry out lab work, so people tend to first carry out a pilot study and look for variation in a few individuals, then assay those same variants in the complete sample of individuals and populations. And since most such investigations are carried out by people of European ancestry, most of the time the variants of interest were detected in samples of European ancestry. In addition, some of the known variants for classical markers and RFLPs came from disease studies, which (not surprisingly) have also tended to focus on populations of European ancestry.

This procedure, therefore, introduces an **ascertainment bias**: since the variants of interest were first detected in populations of European ancestry (and often studied precisely because they exhibited significant variation in such populations), the deck has been artificially stacked to detect more variation in Europeans than would be revealed by studying loci selected at random. The result is that estimates of diversity are inflated relative to non-European populations—which is precisely what is seen in Table 10.1. The fact that this is the most likely explanation can be seen in the contrasting results based on mtDNA and nuclear DNA sequences, where no such ascertainment bias exists because both mtDNA and the nuclear DNA sequences were chosen for sequencing and analysis without any preexisting knowledge concerning the patterns of variation. And Africans have the highest diversity values for both mtDNA and nuclear DNA, which (as we shall in Chapter 14) is in keeping with the hypothesis that modern humans arose in Africa.

What about the STR loci? These were also selected for analysis based on knowing that they were variable in other populations of predominantly European ancestry and hence were also subject to the same sort of ascertainment bias as the classical markers and RFLPs. So why don't the STRs also show the highest diversity in Europeans? Presumably, it is because the STR loci are so variable, with lots and lots of alleles per locus, that this high level of variation compensates for the ascertainment bias—an STR locus that is variable in Europeans is likely to also be variable in non-Europeans, whereas the same is not true for a classical marker or an RFLP. In keeping with this explanation, note that the H values for the STRs are 2–3 times higher than the H values for the classical markers and RFLPs. Also note that while Africans have the most diversity for STRs and mtDNA and nuclear DNA sequences, Europeans have the next highest diversity for STRs whereas Asians have the next highest diversity for mtDNA and nuclear DNA sequences—and this is likely to reflect the ascertainment bias in how the STRs were chosen for analysis.

Ascertainment bias is thus an extremely important issue that—as can be seen in Table 10.1—can profoundly influence the outcome of all sorts of genetic analyses, and ascertainment bias remains a serious issue. The SNP chips that are commonly used to generate genome-wide SNP data, for example, all suffer from ascertainment bias in how the SNPs to be genotyped were chosen for inclusion on the chip. Often, the ascertainment is poorly described (if it is described at all). Fortunately, there are ways that ascertainment bias can be minimized—for example, new SNP chips have been devised in which the ascertainment used to choose the SNPs that are genotyped on the chips is clearly described (e.g., Patterson et al. 2012). Armed with such knowledge, there are ways to try to account for the ascertainment bias in the subsequent analyses. Moreover, with complete genome sequencing all variants are detected without any ascertainment bias, so as the field in general moves more and more toward complete genome sequencing, it can be expected that ascertainment bias—at least as it applies toward choosing which polymorphisms to genotype—will no longer be a concern.

GENETIC DISTANCES BETWEEN POPULATIONS

The next issue to discuss is that of genetic distances between populations. There are an enormous variety of methods for estimating genetic distances between populations, too many to discuss in detail. Here, we will go over a general method that is widely used and applicable to any kind of genetic data and in the next chapter cover some methods specifically for DNA sequence data. The general method is based on partitioning the total genetic variance into within-population and between-population components. The idea is that if we have allele frequencies of one sort or another from at least two groups, then we can calculate the total genetic diversity (i.e., putting everyone into one population) and then examine our two groups to figure out how much of the total genetic diversity is due to diversity within groups and how much is due to diversity between groups. The bigger the diversity between groups (relative to diversity within groups), the bigger the genetic distance between the groups. Formally, if we define H_T as the total genetic diversity (obtained by assuming that all of our samples come from one group and calculating H as defined previously) and H_S as the average genetic diversity within groups (which is obtained by calculating H for each individual group and then taking the average value across the groups), then if we subtract the average within-group diversity from the total diversity, and divide by the total diversity, we get a measure of genetic distance called F_{ST}:

$$F_{ST} = (H_T - H_S)/H_T$$

The name "F_{ST}" comes about for historical reasons, as it turns out that this formulation is related to "F-statistics" that were derived by Sewall Wright in the 1920s to express the probability of drawing alleles that are identical by descent in individuals from the same versus different populations (it is also no coincidence that back in Chapter 5 when we defined the inbreeding coefficient by this probability, we designated it F).

Let's consider a few examples to see how the equation works. Suppose I genotype a SNP with two alleles (call them A and B) in each of two populations, and that in both populations the frequencies of A and B are

both 0.5. Let's also suppose that our sample sizes are big enough that we can ignore the correction for sample size. Then $H_1 = H_2 = 1 - (0.5)^2 - (0.5)^2 = 0.5$, and H_S is the average of H_1 and H_2, which is also 0.5. And in the total sample, the allele frequencies are also 0.5, so H_T is also 0.5, and, therefore, F_{ST} is (0.5−0.5)/0.5 = 0. Hopefully, this makes sense—when the allele frequencies are the same in two populations, then the genetic distance is zero because the two populations do not differ genetically. Moreover, note that all of the variation is due to within-population variation and not to any genetic differences between populations.

Now suppose that in the first population, the frequency of $A = 1$ and $B = 0$, while in the second population, the frequency of $A = 0$ and $B = 1$. Again, assuming that A and B have sample sizes that are the same and large enough that we can ignore the correction for sample size, then in the combined sample, we have the frequency of $A = 0.5$ and the frequency of $B = 0.5$, so $H_T = 0.5$. And both H_1 and H_2 are zero, so $H_S = 0$, and, therefore, $F_{ST} = 1$. And hopefully this also makes sense—when two populations are fixed for different alleles, then they are as different genetically as can be. Moreover, note that in this situation, all of the genetic variation is due to the differences between individuals from different populations—individuals from the same population have identical genotypes, so there is no within-population variation. So, F_{ST} values can range from 0 (meaning no genetic difference) to 1 (maximum genetic difference).

As a measure of genetic distance, F_{ST} values are quite flexible and hence quite useful. They can be computed either for pairs of populations (as above) or for many populations at once, and they can be computed for single genetic loci or averaged over many loci. They can be computed and thus directly compared for various types of genetic markers and for unequal sample sizes (using the appropriate correction for sample size). They have a natural and straightforward interpretation, as they reflect the proportion of the total genetic variance that is due to differences among populations. Moreover, they can easily be extended to additional hierarchical subdivisions of the total genetic variance—for example, if one has sampled several populations from each of several continents, one can compute how much of the total genetic variance is due to differences among populations from different continents, how much is due to differences among populations from the same continent, and how much is due to differences between individuals from the same population.

Still, F_{ST} values are not without drawbacks. For example, the maximum F_{ST} value of 1 is obtained only when populations are fixed for different alleles; if two populations are each polymorphic for different alleles at a locus, then the F_{ST} value will be less than 1 (try an example for yourself if you need convincing that this is true—for example, if a locus has four alleles and population 1 has alleles A and B each at a frequency of 0.5, while population 2 has alleles C and D each at a frequency of 0.5, then F_{ST} is 0.33, even though the two populations do not have any alleles in common). Also, even though by definition F_{ST} values are supposed to be between 0 and 1, with the correction for sample size, it is possible to get negative F_{ST} values. Although how to interpret negative F_{ST} values is a source of consternation for some, they can simply be viewed as a possible result when one is trying to estimate an F_{ST} value close to zero—if the true value is zero, then by chance your estimate may be slightly less, and hence negative—and so the practical thing to do is to set negative F_{ST} values equal to zero. Still, because of the aforementioned (and other) issues, several modifications and extensions of the simple formulation of F_{ST} defined previously have been proposed—too many to go into detail here. For our purposes, the simple version of F_{ST} defined previously will suffice, and that is what we will use unless explicitly stated otherwise.

Let's see what some F_{ST} values look like for humans. Table 10.2 provides F_{ST} values for several different kinds of markers in worldwide populations. Note that all of these F_{ST} values are remarkably similar and indicate that about 8–14% of the genetic variation in humans is found between populations, so about 86–92% of the variation is found within populations. Don't be confused by the fact that overall levels of variability are quite different for these different markers (as was seen in Table 10.1); recall that F_{ST} is the proportion of the total genetic variance that can be attributed to differences between populations. So, whether the actual amount of genetic variation is large or small for a particular genetic marker doesn't matter, what matters is how much of that variation is due to differences between populations.

If it hasn't already occurred to you, you should also be wondering about the populations that were

TABLE 10.2 ■ F_{ST} values for worldwide populations, based on different types of genetic markers

Type of marker	F_{ST}
Classical markers[a]	0.12
RFLPs[b]	0.14
Alu insertion polymorphisms[c]	0.13
Autosomal STRs[d]	0.09
Genome sequences[e]	0.08

[a] Cavalli-Sforza et al. (1994).
[b] Bowcock et al. (1987).
[c] Stoneking et al. (1997).
[d] Pérez-Lezaun et al. (1997).
[e] 1000 Genomes Project Consortium (2010).

included in these estimates—surely all of these markers weren't genotyped in the same populations, so how does choice of populations influence these estimates? It turns out that on a global scale, you will get basically the same estimates as long as you have a few populations each of European, Asian, and African ancestry, regardless of which populations you actually choose. So, the overall F_{ST} value of around 0.08–0.14 is robust to the choice of populations, as long has you have a somewhat globally diverse set of populations.

And what this F_{ST} value literally means is that if you take any single human population from around the world, you will find that it has on average about 82–89% of the total genetic variation that is present in the entire human species. By contrast, different populations of chimpanzees have F_{ST} values ranging from 0.09–0.32, while the F_{ST} value between Bornean and Sumatran orangutans is 0.28 (Fischer et al. 2006). So, human populations are genetically quite closely related, more so than for populations of other apes—although the F_{ST} values in humans are significantly bigger than zero, indicating that the human species as a whole is not a single randomly mating population. This should not come as a surprise to you—if humans were mating at random, then most of us should be choosing someone from China or India as mates, since together they comprise more than 35% of the world's population! In fact, the chances that you will mate with someone from a nearby geographic location are vastly greater than the chances that you will mate with someone from a distant location (although this is changing with increasing globalization), and indeed as we shall see in Chapter 14, there is a very strong geographic component to the genetic variation in humans. Still, the relatively low F_{ST} value for humans has implications for whether or not there is any biological basis to the concept of human "races," also discussed in Chapter 14. Moreover, even though the average F_{ST} value is relatively low for human populations, some genes have unusually high F_{ST} values, which as discussed in Chapter 18 may be an indication of recent positive selection; the Y chromosome also has an elevated F_{ST} value for reasons discussed in Chapter 19 (so, lots to look forward to!).

Analysis of Molecular Variance (AMOVA) and Mantel Tests

One of the virtues of F_{ST} values mentioned previously is that they are readily interpreted as the between-population component of the total genetic variance, and they can be easily extended to additional hierarchical levels. In the early 1990s, Laurent Excoffier and colleagues provided a useful statistical framework for analyzing and decomposing the total genetic variance—analogous to the ANOVA (analysis of variance) that is a standard tool in statistics—that they called **AMOVA** (for analysis of molecular variance; Excoffier et al. 1992). Analysis of molecular variance differs from the standard ANOVA in two important attributes. First, it allows the incorporation of the molecular differences in the alleles into the calculation of the variance components. That is, when we previously calculated F_{ST} values, we considered only the allele frequencies and ignored any information as to how different the alleles were at the molecular level (e.g., number of substitutions for DNA sequences, number of repeat differences for STR alleles, etc.). The F_{ST} value for alleles that differ by one nucleotide substitution is the same as that for alleles that differ by 10 nucleotide substitutions, as long as the allele frequencies in the former case are the same as in the latter case. But AMOVA can incorporate the molecular differences between alleles (as well as the allele frequency differences) into the variance components, and to distinguish such "molecular F_{ST}" values from traditional F_{ST} values (that are based solely on allele frequency differences), these are designated "Φ-statistics" (pronounced "phi statistics" since Φ is the Greek letter Phi). So, the analogue of F_{ST} that incorporates the molecular differences between alleles is Φ_{ST}. You may also come across R_{ST}, which is a version of F_{ST} that is designed specifically for STR loci and incorporates the difference in the number of repeats between alleles at each STR locus. Under a stepwise mutation model that is generally considered appropriate for STR loci, alleles that differ by two repeats are considered more different than alleles that differ by one repeat. The take-home message is that molecular F_{ST} values such as Φ_{ST} and R_{ST} often provide more accurate measures of genetic distance than the simple version of F_{ST} that is based solely on allele frequency differences.

The second important attribute of AMOVA is that the statistical significance of the variance components is determined by a **permutation test**. In a standard ANOVA, the statistical significance of variance components (i.e., whether or not the between-group variance is significantly bigger than zero) is conventionally assessed by tests based on the normal distribution. This is all well and good if your underlying data do in fact follow a normal distribution (i.e., the familiar bell-shaped curve), but if they don't, then you can easily come to the wrong conclusions. With molecular genetic data—that is, the frequency of different alleles, sequences, or haplotypes—we usually have no clue as to whether or not a normal distribution is a reasonable assumption. The permutation test nicely gets around this issue, because it makes no assumption about the underlying distribution of the data.

It is easiest to explain the permutation test by way of example. Suppose we have two urns, one with 40 white balls and 60 black balls, and one with 60 white

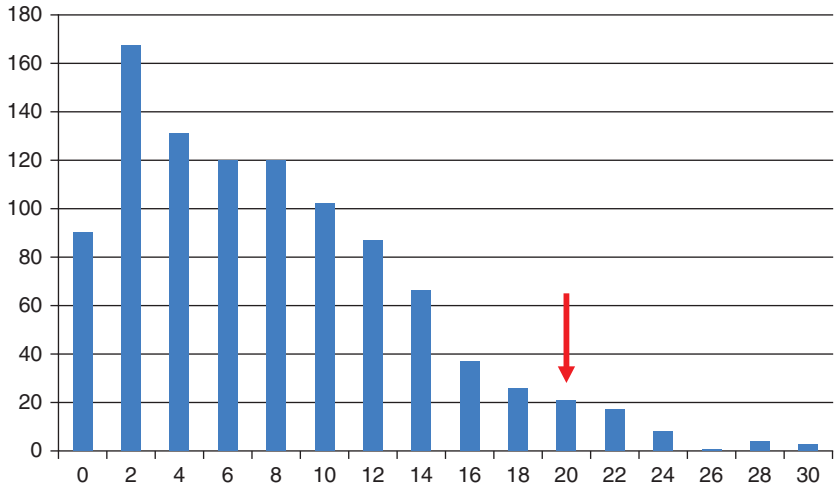


FIGURE 10.2

Bar graph of the outcomes of 1000 permutation tests for the example described in the text. The X-axis is the difference in the number of white balls between the two urns, and the Y-axis is the number of times that difference was observed out of 1000 replicates. The arrow indicates the observed value.

balls and 40 black balls. We want to know whether the distribution of white and black balls differs significantly between these two urns. First, calculate some statistic that measures what we want to know—to keep it simple, let's just take the difference in the number of white balls by subtracting the smaller value from the larger, so in this case, the observed difference is 20. Now, dump out all the balls, then randomly choose 100 balls and put them into one urn, put the remaining 100 balls into the other urn, and calculate the difference in the number of white balls—this is one outcome from randomly permuting the total sample of balls into two subsamples of the same size as in our case. Repeat this procedure 1000 times, and you'll get a distribution of the expected outcomes based on random permutations of the balls into two urns. Figure 10.2 shows the results I got when I did this (although I confess I used a computer to simulate the permutations and generate the figure, rather than actually counting balls!). To figure out what the chances are of obtaining our observed value of 20 for the difference in the number of white balls in the two urns, count the number of permutations that gave a value of 20 or more. Why 20 or more, and not just exactly 20? Because if a large number of outcomes are possible, any single outcome may have a low probability associated with it—flip a coin 1000 times, and the chance that you will get exactly 500 heads and 500 tails is only about 2.5%, even though that is the most likely outcome. And since we usually want to know how extreme our observation is, compared to the expected distribution (based on permutations), we therefore want to know how often we would observe our actual value plus anything even more extreme. In the permutations in Figure 10.2, there were 54 occurrences of a difference of 20 or more out of 1000 permutations, so the empirical probability (or p value) of our observation is 54/1000 = 0.054. So what do we conclude about our two urns—does the distribution of white and black balls differ significantly between the two urns? Convention says that the p value has to be less than 0.05 to be called significant, but if you recall our discussion about significance testing back in Chapter 4, rather than stating that the difference is significant or not, we report the p value and leave it up to the reader to decide whether, in the context of the analysis and the importance of the outcome, a p value of 0.054 is small enough to conclude that the difference between the two urns is unlikely to have arisen by chance alone.

You may think that permutation testing is a rather complex way to answer a simple question, especially when there are straightforward statistical tests that one could apply, and in this particular case you would be correct. But suppose we had four urns: one with 43 red balls, 23 blue balls, 16 green balls, 10 black balls, and 4 white balls; another with 27 red balls, 19 blue balls, 12 green balls, 4 black balls, and 7 white balls; another with 31 red balls, 22 blue balls, 13 green balls, 7 black balls, and no white balls; and the last with 132 red balls, 54 blue balls, 48 green balls, 27 black balls, and 20 white balls. By far, the easiest way to determine how different these are from one another would be by a permutation test. Or, to use a more pertinent example, suppose we want to assess the overall differentiation among mtDNA sequences from several populations, each with a different sample size, some sequences shared both within and between populations, some sequences shared only within populations,

FIGURE 10.3

Map of the Caucasus region, showing the four major language groups spoken in the region: North Caucasian, South Caucasian, Indo-European, and Turkic.

and some sequences found only in a single individual, while taking into account the number of nucleotide substitutions between sequences. An AMOVA with permutation tests is really the only way to go.

The third useful feature of AMOVA is that Laurent Excoffier and his group have gone to a good deal of time and trouble to provide an extremely useful and (relatively) easy-to-use package of programs to carry out AMOVA (and related analyses) called ARLEQUIN. This is not a trivial issue; there are many methods that would see much wider use if it wasn't for the fact that there is no readily available software to carry out the method, or the software is poorly designed or hard for the average user to implement.

Anyway, let's go through an AMOVA of genetic data to see how to interpret the results. Figure 10.3 is a map of the language diversity in the Caucasus region. As you can see, most populations speak either North Caucasian or South Caucasian languages. However, Armenians speak an Indo-European language, while Azerbaijanis speak a Turkic language. We, therefore, would like to know the genetic relationships of Armenians and Azerbaijanis. Are they more closely related genetically to their linguistic neighbors (other Indo-European–speaking groups and other Turkic-speaking groups, respectively) or to their geographic neighbors (other groups from the Caucasus who speak Caucasian languages)? While there are a variety of analyses we could do to investigate this question, let's see what AMOVA tells us. My late colleague Vano Nasidze, who was born in the Republic of Georgia in the Caucasus, collected samples throughout the Caucasus region during the 1980s (when it was a much safer region to travel in than it is now). He later analyzed mtDNA variation (sequences of the first hypervariable segment of the mtDNA control region, or HV1) in these samples in my laboratory, along with HV1 sequences from European populations (speaking Indo-European languages) and Western Asian populations (speaking Turkic languages). If we first group populations on linguistic

criteria, we have groups that speak Caucasian, Indo-European, and Turkic languages, with Armenians included with other Indo-European–speaking groups and Azerbaijanis included with other Turkic-speaking groups. We can then ask how much of the total genetic variance in HV1 sequences is due to differences among individuals from within the same population (this is the *within population* variance), how much is due to differences among populations speaking languages from the same family (this is *among populations from the same group* variance), and how much is due to differences among populations speaking languages from different families (this is *among group* variance). Here are the AMOVA results (Nasidze et al. 2004):

Within population: 96.8%
Among populations, same language group: 2.3%
Among language groups: 0.9%

Note that by far the biggest component of the genetic variance is accounted for by differences among individuals from the same population—only 3.2% of the genetic variance is due to differences among populations. So, again we see that there is very little genetic differentiation among human populations. But of this 3.2% that reflects genetic variance among populations, most of it is due to variance among populations from the same language group. If our language classification corresponded to the genetic structure of these groups, then populations speaking languages from the same family should be more similar genetically than populations speaking languages from different families—but they aren't. So, the language classification is a poor fit to the genetic structure of these populations.

Let's see what happens instead with a geographic classification of populations. If we classify populations as being European, West Asian, or Caucasian (from the Caucasus), then the only difference between this classification and the aforementioned linguistic classification is that Armenians are classified as Caucasian instead of as Indo-European, and similarly Azerbaijanis are classified as Caucasian instead of as Turkic-speaking. Otherwise, all of the other Indo-European populations go into the Europe group, and all of the other Turkic-speaking populations go into the West Asian group. Now when we do the AMOVA, we get the following results:

Within population: 95.8%
Among populations, same geographic group: 1.7%
Among geographic groups: 2.5%

Note that the within-population variance is slightly (and nonsignificantly) different than in the previous analysis; this often happens because of the way the variance components are calculated by the computer program (Arlequin) that was used to carry out the AMOVA. More importantly, with this classification the variance among populations from the same geographic group is lower than the variance among populations from different geographic groups. Thus, populations from the same geographic region are more similar genetically than populations from different geographic regions, and a geographic classification provides a better fit to the data than a linguistic classification. The implication of these results is that even though Armenians speak an Indo-European language, genetically they are more similar to other populations from the Caucasus than they are to other Indo-European–speaking populations. The same holds for Azerbaijanis: genetically they are also more similar to other populations from the Caucasus than they are to other Turkic-speaking populations. Other ways of analyzing the genetic data support this conclusion, as do other kinds of genetic data.

How might we explain this discrepancy between the linguistic and genetic relationships of Armenians and Azerbaijanis? The most likely explanation is that the Armenian and Azerbaijani languages were introduced by **language replacement**: in the past, ancestors of Armenians/Azerbaijanis probably spoke a Caucasian language, but then they switched to an Indo-European/Turkic language, respectively, without a significant genetic contribution from Indo-European/Turkic populations. Hence, genetically Armenians and Azerbaijanis are more closely related to their geographic neighbors than to their linguistic neighbors.

This example illustrates how we can use AMOVA to investigate the correspondence between different ways of classifying the populations under study and their genetic structure. Keep in mind that this is not a formal test—in this case, we don't know whether the geographic classification is actually significantly better than the linguistic classification. But as an exploratory tool for evaluating which of several different classifications provides the best fit to the genetic structure of the populations of interest, AMOVA is extremely valuable and versatile. There is also an extension of AMOVA, called SAMOVA (for Spatial Analysis of Molecular Variance) that tries to come up with the overall grouping of populations that provides the best fit to their genetic structure. Often, the best grouping can also be identified from visual displays of the genetic relationships of the populations (using methods described later), but SAMOVA still provides an independent check on the presumed best classification.

A related analysis that we frequently want to do is to compare the genetic distances between each pair of populations in our analysis to some other distance measure between them. Usually, these are geographic

Genetic	pop1	pop2	pop3	pop4
pop1		A	B	C
pop2	A		D	E
pop3	B	D		F
pop4	C	E	F	
Geographic	pop1	pop2	pop3	pop4
pop1		a	b	c
pop2	a		d	e
pop3	b	d		f
pop4	c	e	f	
Permute 1	pop1	pop2	pop3	pop4
pop1		d	e	f
pop2	d		b	c
pop3	e	b		a
pop4	f	c	a	

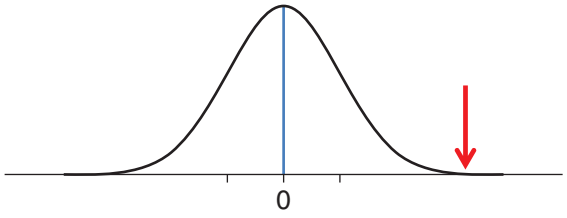


FIGURE 10.4

Rationale behind the Mantel test for a significant correlation between two matrices. Left, suppose we have the genetic distances (top) and geographic distances (middle) for a set of populations. The Mantel test then permutes one matrix randomly (bottom, for geographic distances—the genetic distance matrix is not manipulated), calculates the correlation coefficient with the other matrix, then repeats this process to generate a distribution of correlation coefficients that represent random expectation (right). If our observed correlation coefficient (red arrow) is bigger than most of the randomly generated correlation coefficients, then the observed correlation coefficient is deemed statistically significant.

distances, but these can also be linguistic distances or anything else where we can quantify the differences among populations. For the sake of example, let's assume that we have a matrix of genetic distances among a set of populations, and we want to know whether they are correlated with the geographic distances among these populations. The straightforward way to do this would be to calculate the correlation coefficient between the genetic distances and the geographic distances for each pair of populations. As we saw back in Chapter 5 when discussing assortative mating, correlation coefficients can, in theory, range from −1 to 1, where significant negative values indicate a negative relationship (in this case, as geographic distances increase, genetic distances decrease among populations), significant positive values indicate a positive relationship (i.e., as geographic distances increase, genetic distances also increase among populations), and values that do not differ significantly from 0 indicate no relationship. How can we tell if a particular correlation coefficient is significantly different from zero? The usual way is to carry out a statistical test based on the normal distribution, but note that our data violate an important assumption of such tests, namely, that the data points (observations) are independent. That is, conventional statistical tests assume that changing one observation won't influence any of the other observations in the data. But this is not true for our matrix of pairwise distances—if we change one of the populations, we change all of the distances between that population and all of the other populations in the analysis. Hence, the observations in a pairwise distance matrix are not independent, and therefore we cannot carry out a conventional test for the significance of the correlation coefficient.

What can we do? It turns out that there is a convenient procedure based on permutations that neatly answers our question. As depicted in Figure 10.4, what you can do is to take one of the distance matrices and keep it as it is while taking the other distance matrix and randomly exchanging (i.e., permuting) the rows and columns of the matrix. You then end up with a matrix that has the same observations but in a jumbled order. Now calculate the correlation coefficient and then do another random permutation of the distance matrix. Do this 1000 times and you have a distribution of correlation coefficients based on random permutations of your observed data. If your observed correlation coefficient is sufficiently bigger than the majority of the permuted correlation coefficients (by convention, bigger than 95% of them), then you can conclude that the observed correlation coefficient is indeed unlikely to have arisen simply by chance. This procedure, which is another nice example of the power of permutation tests, is known as the Mantel test, after the person who invented it, Nathan Mantel (Mantel

1967). Nathan Mantel was a noted biostatistician who made numerous important contributions—quite fittingly for a statistician, the epitaph on his gravestone reads "one in a million."

DISPLAYING GENETIC DISTANCE DATA: TREES

One of the results from genetic analyses of populations is a matrix of genetic distances (such as F_{ST} or Φ_{ST}) between each pair of populations. These tables can readily reach gargantuan proportions (e.g., for 50 populations there would be 1225 distance values, and for 100 populations there would be nearly 5000 distance values), and while there are mathematically gifted individuals who can look at such tables and see patterns in them, for the rest of us mere mortals we need some other way to identify the relationships in such data. Moreover, humans tend to be much better at seeing things from figures rather than from tables of numbers, so for these reasons methods have been developed for visually displaying the relationships among populations. In this section we consider **trees**, while in the following section we will discuss other methods for identifying and displaying the most important patterns in genetic data from populations.

A tree is simply a branching diagram that depicts the relationships among a set of sequences, haplotypes, populations, species, and so forth. We need a generic label for the things we want to relate in a tree—sequences, haplotypes, populations, species, and so forth—and so for this purpose, we'll borrow a term from systematics and use **OTU** (operational taxonomic unit) when the specific things we want to relate are not important. Since in this chapter we are focusing on analyses where the population (or, in some cases, the species) is the unit of analysis, in this section we will similarly focus on methods specifically for building trees where the OTUs are different populations or species. Trees can also be constructed for individual DNA sequences, haplotypes (e.g., based on Y chromosome SNPs and/or STRs), or genotypes (with sufficiently dense multilocus data), and such trees will be discussed in the next chapter.

There are many ways of constructing trees from genetic distance data—too many to discuss each of them—so we'll consider a few representative methods that are the most widely used in molecular anthropology. The easiest of these to understand goes by the unwieldy name of unweighted-pair-group-method-of-averaging, or **UPGMA** for short. An example of how to construct a UPGMA tree is provided in Box 10.1; the basic idea is to start with the two OTUs with the smallest genetic distance, link them together in the tree, and replace them in the distance matrix with their ancestor, with new genetic distances between the remaining OTUs and this ancestor obtained by averaging all of the relevant genetic distances. This procedure is repeated until all of the OTUs have been added to the tree.

Unweighted-pair-group-method-of-averaging is very quick and easy, even with very large numbers of OTUs, and numerous programs are available to construct UPGMA trees from genetic distance matrices. Hence, UPGMA is a popular method for constructing trees. However, it is important to keep in mind that UPGMA does assume a constant rate of change in the genetic distances (this assumption is behind the averaging that goes on when genetic distances to ancestors in the tree are calculated—see the example in Box 10.1). If the OTUs are sequences, or genetic distances based on mutational differences, then this may be a reasonable assumption, as this corresponds to a molecular clock (and as we shall see in Chapter 12, we can test whether our data are consistent with a molecular clock). But if the OTUs are populations for which we have calculated genetic distances based on allele frequencies, then the UPGMA tree is valid only if the rate of change in allele frequencies over time has been the same in all of the populations. This, in turn, means that all of the processes that can influence allele frequencies, such as genetic drift, inbreeding, changes in population size, migration, and so forth, have been the same in all of the populations—a dubious assumption at best.

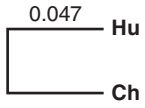
Fortunately, there are other methods for constructing trees from distance matrices that do not assume a constant rate of change. The most widely used of these is the **neighbor-joining** (NJ) method, developed by population geneticist Masatoshi Nei and his student Naruya Saitou (Saitou and Nei 1987)—and the correspondence between the name of the developer and the first three letters of the name of this method probably isn't a coincidence! This method works somewhat differently than UPGMA, in that NJ starts with a starlike tree with all OTUs connected to a single node and then assumes that a pair of OTUs should be connected by an ancestral node (Figure 10.5)—that is, joined as neighbors. To figure out which pair of OTUs should be joined, you start by using the genetic distances to estimate the length of the entire tree when each pair of OTUs is joined. Don't be concerned about how to do this—the equations have been worked out and they are too complex for us to worry about, and anyway nobody does this by hand when there are easy-to-use computer programs that are freely available to construct NJ trees. Take the pair of OTUs that give the overall smallest tree length when they are joined, join them, and then calculate a new matrix of the tree lengths that are obtained when each pair of OTUs is joined. In this new matrix, the OTUs joined in the previous step are replaced by their ancestral node, so now the number of OTUs in the matrix is reduced by one.

BOX 10.1 ■ Example of UPGMA Tree

Consider the following matrix of genetic distances among humans (Hu), chimpanzees (Ch), gorillas (Go), orangutans (Or), and gibbons (Gi):

	Hu	Ch	Go	Or	Gi
Hu	XXX				
Ch	0.094	XXX			
Go	0.111	0.115	XXX		
Or	0.180	0.194	0.188	XXX	
Gi	0.207	0.218	0.218	0.216	XXX

We first take the pair of OTUs with the smallest distances, in this case Hu and Ch, and link them together:



Note that we get the distances along the branches leading to Hu and Ch by assigning half the genetic distance between them to each branch (i.e., we divide their genetic distance by 2). We then replace Hu and Ch in the distance matrix by their grouping and get new genetic distances by taking the average of each other OTU with Hu and Ch:

Go vs. Hu-Ch = (0.111 + 0.115)/2 = 0.113

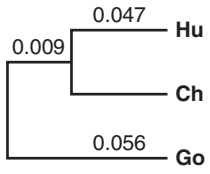
Or vs. Hu-Ch = (0.180 + 0.194)/2 = 0.187

Gi vs. Hu-Ch = (0.207 + 0.218)/2 = 0.212

This gives us the following distance matrix:

	Hu-Ch	Go	Or	Gi
Hu-Ch	XXX			
Go	0.113	XXX		
Or	0.187	0.188	XXX	
Gi	0.212	0.218	0.216	XXX

The smallest distance now involves Go vs. Hu-Ch, so we link Go to the tree and obtain branch lengths by assigning half the Go vs. Hu-Ch distance to each lineage:

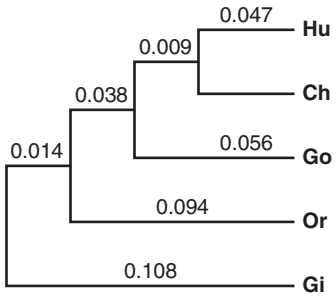


In case it isn't clear, the length between the Hu-Ch node and the Hu-Ch-Go node of 0.009 is obtained by subtracting the length for the Hu and Ch lineages of 0.047 from the length for the Go lineage of 0.056.

As before, we replace Go and Hu-Ch by the new group Hu-Ch-Go and obtain new genetic distances from the remaining OTUs by averaging their distances to Hu, Ch, and Go, which then gives the following distance matrix:

	Hu-Ch-Go	Or	Gi
Hu-Ch-Go	XXX		
Or	0.187	XXX	
Gi	0.214	0.216	XXX

The smallest distance involves Or with Hu-Ch-Go, so we make this link, and then the last to be added is Gi, resulting in the final UPGMA tree:



Voilà!

Keep doing this until all OTUs have been joined—the result is the NJ tree.

Neighbor-joining trees are thus **minimum evolution** trees, in that the criteria used to obtain the NJ tree is that the overall distances along the tree are minimized. However, keep in mind that the algorithm used to construct NJ trees, which finds minimal trees in a stepwise fashion by proceeding with the shortest tree obtained at each step, is not guaranteed to find the actual minimum evolution tree. That is, the tree with the overall shortest distance for all of the OTUs (the minimum evolution tree) may differ from what you will get when you take the shortest trees at each step. Still, NJ has been shown to perform very well with simulated data (so one knows what the real answer is), is quite fast and efficient even for very large data sets (with hundreds or even thousands of OTUs), and is justifiably quite popular for constructing trees from genetic distances.

While NJ and UPGMA trees provide clear indications of population relationships, in that populations in the same **clade** or branch are more similar genetically

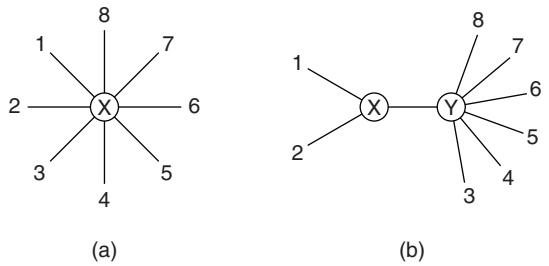


FIGURE 10.5

Neighbor-joining method of tree construction. The procedure starts by assuming a starlike phylogeny for all of the taxa; each pair of taxa is then joined together and the resulting sum of the branch lengths across the tree computed. The pair of taxa that result in the tree with the smallest length are then joined as neighbors (in this case, 1 and 2 are joined), the number of taxa is reduced by one (1 and 2 are replaced by the new node X), the distances recomputed, and then the process repeated until the tree is complete. Modified with permission from Saitou, N., and Nei, M., "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution* 4:406, 1987.

than populations in different clades, they don't in and of themselves provide any indication of how strongly the data support the relationships in the tree. A convenient way of doing this is by **bootstrap analysis**, which is another form of a permutation test. Suppose you have a set of populations for which you have genotyped 50 loci, calculated average pairwise F_{ST} distances across all 50 loci, and then constructed an NJ tree. To do the bootstrap analysis, sample 50 loci from your data set at random but with replacement—so some loci will be present more than once in this bootstrap sample, while other loci won't be present at all. Calculate the genetic distances and the NJ tree and then repeat this 1000 times or so. Count how many times each clade of populations in your original NJ tree is represented in the set of 1000 bootstrap trees and that is the bootstrap support value for that clade. The idea is that if a particular clade is supported by many loci in your data set, then you expect to find that clade in most of your bootstrap trees. But if a particular clade is supported by only a few loci, then that clade won't occur very often among the bootstrap trees—only if one or more of those few loci end up in the bootstrap sample will you find the clade in the resulting tree. In general, for trees based on species, we expect to find high support for the clades, as we expect the relationships among species to be supported by many genetic changes. So it is common to demand that bootstrap values should be around 95% for a particular clade to be considered strongly supported by the data. But for closely related populations (such as human populations), it may very well be the case that a perfectly good relationship in the tree does not receive high bootstrap support, because only a very few loci happen to have genetic changes that support that relationship. For such data, we might be happy with bootstrap values as low as 50%—bootstrap values provide a useful guide as to how strongly the data support a grouping of OTUs, but there are no hard and fast rules as to what an "acceptable" bootstrap value should be.

A third type of tree that you may run across are **maximum likelihood** (ML) trees. Maximum likelihood is a common method used in statistics that involves calculating the likelihood (which is related to, but not identical to, the probability) of the data for various models, and then choosing the model that has the highest likelihood of giving rise to the observed data. For example, if we toss a coin 10 times and get 6 heads and 4 tails, we can try various estimates of the probability of getting a head (e.g., 0.1, 0.2, 0.3, … 0.9) and see which one gives us the highest likelihood of observing 6 heads in 10 coin tosses. It should come as no surprise to you that the ML estimate for this probability is 0.6, because that clearly has the highest chance of resulting in the observed data. So even though you may not know it, many familiar estimates that we use routinely (such as taking the arithmetic average of a sample of observations as the best estimate of the mean of the underlying distribution) turn out to be ML estimates.

Maximum likelihood estimates furthermore have the desirable properties that as the sample size increases, they are both consistent and efficient. The statistical meaning of "consistency" is that as the sample size gets bigger and bigger, the ML estimate (such as the arithmetic average) of a parameter (such as the mean) gets closer and closer to the true value, while the statistical meaning of "efficiency" is that as the sample size gets bigger and bigger, the ML estimate is the closest of all possible estimates to the true value. Furthermore, one can calculate the likelihood of the data for the best model and for alternative models and test whether the likelihood of the best model is significantly higher than the likelihood of a specific alternative model. For example, even though in the aforementioned coin toss example, the ML estimate of the probability of getting a head is 0.6, the likelihood of the data (6 heads out of 10 coin tosses) given this probability is not significantly bigger than the likelihood of the data if the true probability of getting a head is actually 0.5—in other words, based on the evidence at hand, we cannot reject the hypothesis that the coin is a fair one.

While ML trees are most often based on DNA (or protein) sequence data, it is possible to use the ML approach to construct a tree for populations based on allele frequency data. This approach is known as continuous ML and assumes that all changes in allele

frequencies are due to genetic drift—no selection, no migration, and no new mutations. For genetic data in which the mutations are older than the populations, such as most SNP data, the assumption of no new mutations is not an issue. However, it does mean that in principle continuous ML should not be used to construct trees from STR data, where new mutations commonly occur within populations, but this doesn't stop people from doing so!

The big advantage of ML over other approaches such as NJ and UPGMA is that you can use a likelihood ratio test to see whether one tree provides a significantly better fit to the data (actually, to see whether the data are significantly more likely given one tree vs. another). Given all these advantages of ML, you might wonder why anyone should bother using other methods such as NJ or UPGMA to construct trees. Alas, everything comes with a drawback (TAANSTAFL, as you might remember from Chapter 7), and in the case of ML, it is computationally quite intensive to construct ML trees, especially for lots of OTUs. In fact, for complex data sets and/or models, it can prove computationally impossible to come up with the tree in a reasonable amount of time (such as your lifetime!). Still, given the advantages of the ML approach for suitable data (i.e., where the assumptions are reasonable), it should be the method of choice for constructing population trees if it can provide an answer in a reasonable amount of time.

An important aspect of constructing trees is determining where the root lies, which then corresponds to the common ancestor of the OTUs. This aspect will be discussed in detail in the next chapter. For now, keep in mind that an important difference between UPGMA trees and other kinds of trees such as NJ and ML is the fact that the UPGMA method assumes a constant rate of evolution. This means that the root of a UPGMA tree can be placed at the midpoint of the longest path connecting two OTUs in the tree; this is known as **midpoint rooting**, and an example is given in Figure 10.6. By contrast, NJ and ML trees do not assume a constant rate of change, which thus means that the resulting trees are unrooted, and therefore additional information is required to root the tree—how one does this is discussed in the next chapter. One could also use midpoint rooting with an NJ or ML tree, but midpoint rooting does imply a constant rate of change in your genetic distances, and if you are willing to make that assumption, then you might as well use UPGMA instead (because it has been shown that if you do have a constant rate of change in your genetic distances, then UPGMA is in fact the best method to use). However, given the existence of widely available programs that will readily construct NJ (and ML) trees, unless there is good reason to assume a constant rate of change in the genetic distances, there is no excuse not to construct an NJ or ML tree instead of a UPGMA tree. You may wonder if in practice it really makes any difference; in fact, not only is it easy enough to construct artificial examples where UPGMA will give you the wrong answer, there are cases involving real data where UPGMA gave misleading results, and one of these is presented in Figure 10.7. In fact, I still occasionally receive manuscripts to review where authors have blithely constructed a UPGMA tree (usually out of ignorance about this assumption of a constant rate of change) when an NJ (or ML) tree would be preferable, and as you can probably imagine I do not comment favorably on this!

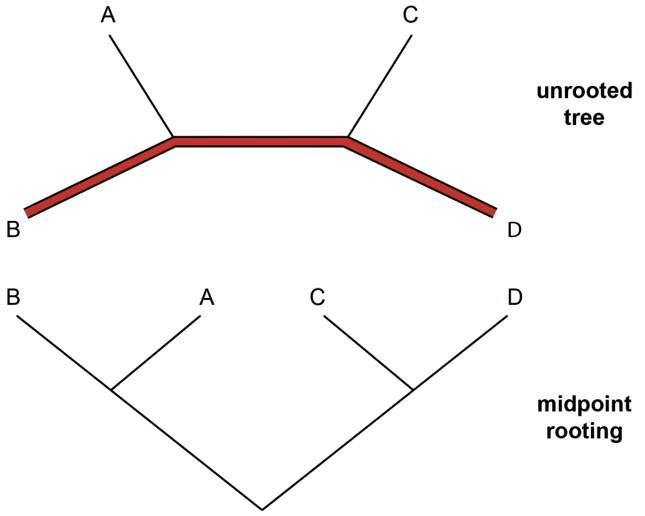


FIGURE 10.6

Midpoint method for rooting a tree. With the unrooted tree at the top, the red bar shows the longest path between any two operational taxonomic units, so the midpoint root (bottom tree) is obtained by placing the root at the middle of this longest path.

One last but very important issue concerning trees is their interpretation. The methods for constructing trees from genetic data on populations have their origins in methods for reconstructing the evolutionary history of different species, and the resulting trees are sometimes called **phylogenetic trees** or **phylogenies**. The branching pattern of trees for species (or other OTUs) that do not exchange genes does indeed reflect the history of the genetic divergence of the species, and a phylogenetic interpretation is perfectly valid. However, populations within a species are by definition not reproductively isolated and hence may very well be exchanging genes via migration. Thus, genetic distances among a set of populations reflect both the history of divergence of populations from common ancestral populations and subsequent migration among populations. Teasing apart the relative importance of these two factors—divergence versus migration—remains an extremely difficult issue.

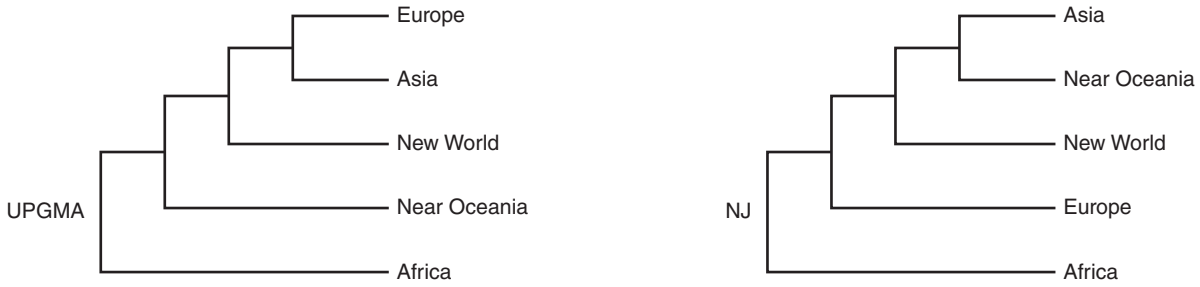


FIGURE 10.7

An example of differences in population relationships for UPGMA versus NJ trees constructed from the same data. Left, the UPGMA tree groups Europe with Asia. Right, the NJ tree, which allows for unequal rates of change along branches of the tree, groups Asia with Near Oceania and the New World. NJ, neighbor-joining; UPGMA, unweighted-pair-group-method-of-averaging. The data used for these figures are modified with permission from Nei M., and Roychoudhury, A., "Evolutionary relationships of human populations on a global scale," *Molecular Biology and Evolution* 10:927, 1993.

To further illustrate this point, there are two very different ways of interpreting F_{ST} values. The population geneticist Sewall Wright showed that if you assume no new mutations and no migration, then F_{ST} increases in proportion to the time of divergence between two populations (Wright 1943):

$$F_{ST} = 1 - e^{(-t/2Ne)}$$

where t is the divergence time and N_e is the effective population size (assumed to be the same for the two populations). F_{ST} values, therefore, become bigger between populations as divergence time increases, because of genetic drift (random changes in allele frequencies over time). So, according to this view, the branching pattern in a tree based on F_{ST} values reflects the history of population divergence events.

But we can just as readily assume a model with no branching history for the populations, and instead populations are simply exchanging migrants every generation. Then under such a model, it can be shown that:

$$F_{ST} = 1/(1 + 4N_e m)$$

where m is the migration rate (Wright 1940). So, the more migration between a pair of populations, the smaller the F_{ST} value between them.

Which view is correct? The cold hard truth is that either could be correct—or a combination of the two—and we have (at present) no good way to distinguish between these two explanations. Thus, trees based on genetic distances among OTUs that are capable of exchanging migrants should be viewed as a visualization of the patterns of overall genetic similarity in the data—no more, no less. Two OTUs that appear to have recently diverged from a common ancestor in a tree could indeed have recently diverged, or they could have diverged further back in the past but appear to be more similar genetically because of migration. Unfortunately, many investigators seem to be under the mistaken impression that just because you can construct a tree for your populations based on phylogenetic methods, the tree therefore shows the divergence history of the populations and migration can be ignored. So, the take-home message is to be wary of how trees depicting the relationships of OTUs that can exchange migrants are interpreted.

DISPLAYING GENETIC DATA: MULTIDIMENSIONAL SCALING, PRINCIPAL COMPONENTS, AND CORRESPONDENCE ANALYSIS

Trees are just one way of displaying the relationships among a set of OTUs based on genetic data. Another class of useful methods consists of ways to simplify the enormous amount of data in a distance matrix or in a table of allele frequencies for a set of populations, while minimizing the loss of the information that such simplification necessarily entails. While there are several such methods, we will focus here on the three methods that are most commonly used with genetic data, namely, multidimensional scaling (MDS), principal components (PC), and correspondence analysis (CA). These are sometimes referred to as "plot" methods, because they all produce a two- or three-dimensional plot of the relationships among the populations.

Let's start with MDS, which uses a pairwise distance matrix as the input. Suppose we have 50 populations in our pairwise genetic distance matrix. If we were to plot our populations in a 50-dimensional space, there would be a perfect fit between the distances between each pair of populations in this 50-dimensional space and our observed genetic distances, and we could then see which populations are close

together in this space. Alas, visualizing relationships in more than three-dimensional space is beyond the capabilities of us mere mortals, so what MDS does is to place the populations in a space consisting of fewer dimensions while maintaining the distance relationships among the populations as closely as possible. Usually one starts with two dimensions, so the result of MDS is a two-dimensional plot in which populations that are close together are interpreted as being more genetically similar than populations that are far apart in the plot. And how does one actually get such an MDS plot from a distance matrix? That's easy: just get your data into the appropriate input format and then run one of the many computer programs available to carry out MDS analysis! Seriously, there is no point in trying to go through the equations as to how to do this, because they are quite complicated and the equations alone don't give you any useful insights into the method. Instead, we focus here on important points to keep in mind when evaluating and interpreting MDS (and other) plots. One feature of such plots has to do with rotating the dimensional axes, and this point is covered in Box 10.2.

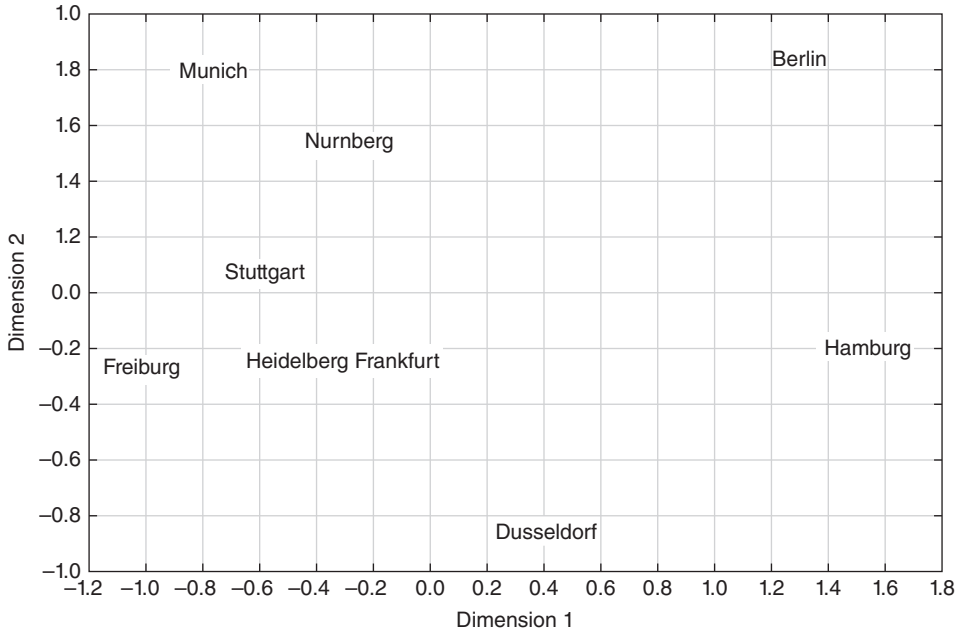
One of the most important aspects of any of these "plot" methods is the evaluation as to how well the resulting plot retains the structure of the data. For

BOX 10.2 ■ An Example of MDS Analysis

Let's start with a table of mileage distances between German cities, which I took from a German road atlas that I happened to have lying around. Here is the matrix of the mileage distances (in kilometers) between each pair of cities:

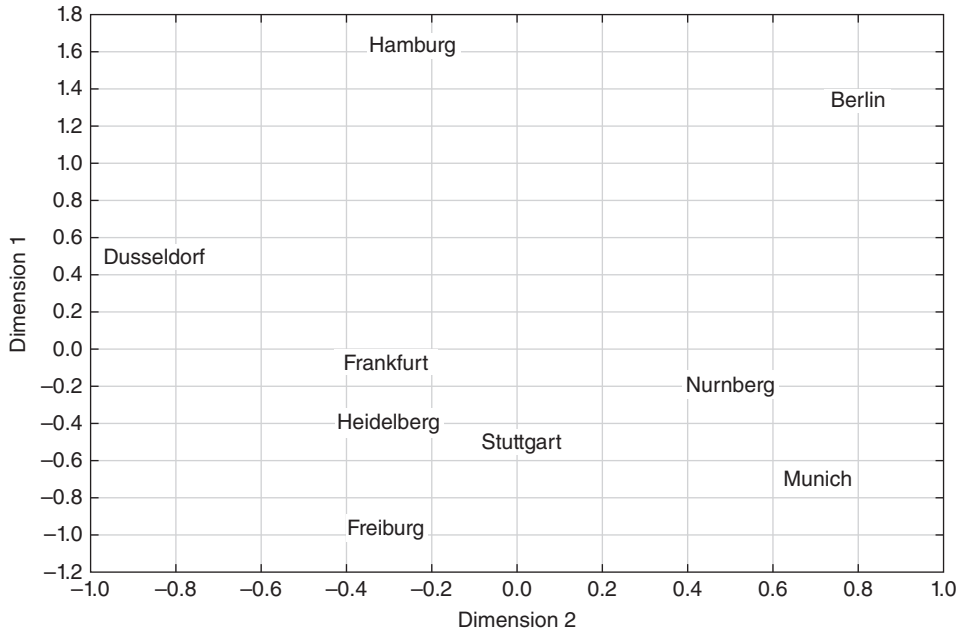
	Berlin	Hamburg	Frankfurt	Dusseldorf	Munich	Freiburg	Stuttgart	Heidelberg	Nurnberg
Berlin	0								
Hamburg	287	0							
Frankfurt	569	494	0						
Dusseldorf	571	428	225	0					
Munich	583	783	393	612	0				
Freiburg	829	756	270	471	345	0			
Stuttgart	624	698	213	412	219	205	0		
Heidelberg	651	578	91	294	321	185	137	0	
Nurnberg	440	613	223	442	165	369	189	260	0

If we now perform MDS on this, using any of the readily available programs to do this, we get the following plot of the first two dimensions:



BOX 10.2 ■ (*Continued*)

As you can readily see, what we get from this is a pretty good fit to a map depicting the spatial relationships of these German cities … wait a minute, actually, this looks nothing like a map of Germany, everything seems to be in the wrong place. What's going on? The important point here is that the values for the dimensions assigned to each data point (German city, in this case) do not have any intrinsic meaning—as long as we maintain the relative relationships among the values assigned to each data point, we are free to manipulate them, however, we see fit. In this case, we can switch the X and Y axes around and plot Dimension 1 on the Y-axis and Dimension 2 on the X-axis. Doing so gives us the following plot:



And now what you see does indeed look very much like a map of Germany—if you aren't convinced, find a map of Germany and compare it to the aforementioned MDS plot. This is to be expected, of course, because the mileage distances between cities should mostly reflect their geographic distances, with some difference that reflects traveling via roads versus traveling as the crow flies. But the take-home message is that we are free to rotate and flip the axes of a plot (e.g., change positive values to negative values and vice versa) however we like, to aid in interpreting the plot, as long as we are careful to maintain the relative relationships among the data points.

MDS, this evaluation is provided by the stress value, which compares the observed distance values between each pair of populations to those obtained from the plot. Low stress values indicate a good fit between the observed distances and those from the MDS analysis, while high stress values indicate a poor fit. Unfortunately, there is no significance test you can do to indicate whether or not the stress value is acceptable, so the (rather arbitrary) convention has arisen in which stress values below 0.15 (or 15%) are deemed acceptable, while stress values above this are unacceptable. And I do mean unacceptable: stress values above 0.15 indicate that the two-dimensional plot so greatly distorts the relationships in the data that you should not base any conclusions on such plots—although one can find published studies where authors either fail to provide stress values or have noted that the stress value is above 0.15 and yet blithely plunge ahead with interpreting the plot anyway. The take-home message: anytime you come across an MDS plot, you should check to see what the stress value is before you accept any conclusions based on the plot.

If you do carry out a two-dimensional MDS analysis and find that the stress value is unacceptably high, what can you do? One way to improve the analysis is to increase the number of dimensions. How can this help? Consider the following example: suppose we have three populations (A, B, and C), and suppose they all have the same genetic distance between them (say, 100). Now, suppose we try to fit a one-dimensional

model to the data: that is, try to arrange all three populations on a straight line. No matter how you try, you won't get a very good fit to the data. You can try putting A and B at a distance of 100, but then if you put C in between them in the middle, then the distance from A to C and from B to C are both 50 instead of 100:

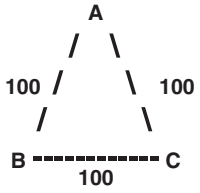
A – – – – – – – – – – C – – – – – – – – – – B
(50) (50)

Or, you could put C at a further distance of 100 from B, and now the distance from A to B and from B to C are both 100, but the distance from A to C is 200:

A – – – – – – – – – – B – – – – – – – – – – C
(100) (100)

No matter what you do, a one-dimensional model simply is inadequate to capture the information in our data—go ahead and try other arrangements if you don't believe me!

But what happens if we instead try to fit a two-dimensional model to these data? In two dimensions, we can place our three populations at the vertices of a triangle with edges all of length 100, and now we have a perfect fit between our two-dimensional model and our data:



So, if a two-dimensional MDS plot results in an unacceptable stress value, try a three-dimension analysis, which then results in a three-dimensional plot (see Figure 10.8 for an example). And if the stress value is still unacceptable, you can increase the dimensions to four (or more), but then it gets more difficult to see what is going on—typically, what you would then do is to make multiple two-dimensional plots of each dimension versus each other dimension.

Principal components and CA are basically variations on the same theme as MDS in that they also take complex multivariate data and simplify the data pictorially. However, an important difference between PC/CA and MDS is that in the former the first axis captures most of the information in the data, and succeeding axes capture less and less of the information, whereas with MDS the axes don't have any inherent meaning, we can rotate and flip them around (see

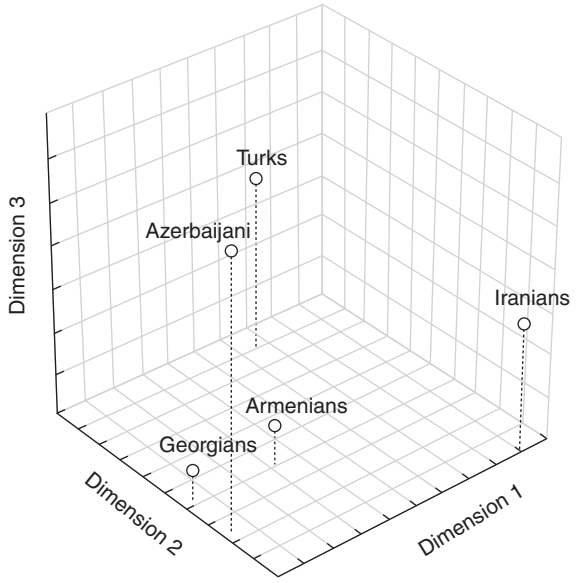


FIGURE 10.8

An example of a three-dimensional MDS plot of population relationships based on complete mtDNA genome sequences. In the first two dimensions, Azerbaijanians and Armenians group with their geographic neighbors (Georgians from the Caucasus), rather than their linguistic neighbors (Turks and Iranians, respectively). However, in the three-dimensional plot, the Turkic-speaking Azerbaijanians are more similar to their linguistic neighbors. Reprinted with permission from Schönberg, A., et al., "High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences," *European Journal of Human Genetics* 19:998, 2011.

Box 10.2 for a demonstration). Principal components usually starts with a matrix of gene frequencies, although it can also be used on distance matrices, while CA requires a matrix of the count of the number of each allele in each population. Correspondence analysis is mainly used when one is interested not only in the relationships among the populations but also among the genes, as these can both be plotted in a CA plot, indicating which gene(s) contribute to where particular populations end up in the plot (see Figure 10.9 for an example). The most common application of CA involves mtDNA and/or NRY haplogroups, as the CA plot then gives an indication of which haplogroups are contributing the most to the relationships of particular populations.

Neither PC nor CA produce a stress value; instead, the amount of the total variance explained by each component is given by the analysis, which then provides a rough idea as to how much of the information

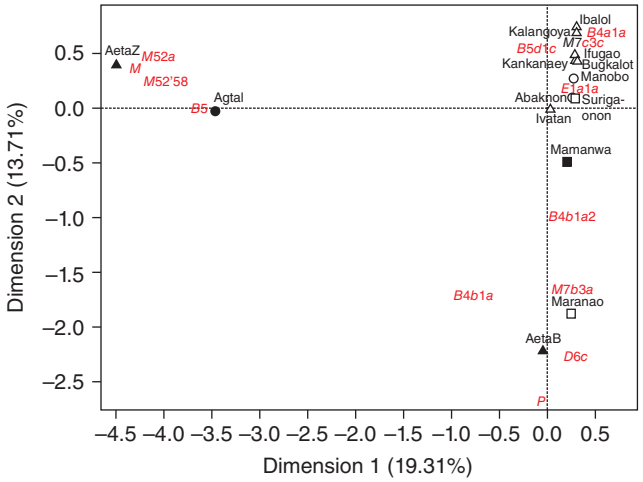


FIGURE 10.9

Example of a correspondence analysis (CA) plot for Filipino ethnolinguistic groups, based on complete mtDNA genome sequences. The symbols indicate various ethnic groups, while the red italic labels indicate mtDNA haplogroups. Thus, the outlier position of the AetaZ and Agtal (upper left of the plot) can be seen to be related to their frequencies of haplogroups M, M52a, M52'58, and B5. Modified with permission from Delfin, F., et al., "Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region," *European Journal of Human Genetics* 22:228, 2014.

in the data is captured by the various components. By definition, the first component will account for the highest proportion of the variance, and each succeeding component will account for less of the variance.

An example of a PC plot is provided in Figure 10.10, along with an NJ tree for the same data. Note that the first PC corresponds closely to the tree in showing a major division between African and non-African populations—in general, the first PC is expected to be similar to a tree, as both capture the most significant information in the data. The second PC, however, indicates that populations from Australia and New Guinea are more similar to the ancestral population than are other populations, and this information is not captured in the tree. Mathematically, it can be shown that PC analysis captures more of the information in the data than tree analysis, and hence some would argue that PC analysis is preferable to constructing trees for population data. My own view is that while PC analysis can indeed be more informative, trees are still useful in that they provide objective evidence for the existence of particular clusters of populations. For example, the tree in Figure 10.10 shows that all of the European populations fall into one branch, or clade. You might be tempted to look at the PC plot in Figure 10.10 and think, well, doesn't this also show a cluster of the European populations? However, our perceptions are influenced by the fact that the dots are colored to correspond to the continental origins of the populations, so we view clusters of dots of the same color as if they were somehow objective clusters in the PC plot. But suppose we hide the information about continental origin—would we still identify a cluster of European populations? Take a look at the version of the PC plot in Figure 10.10 in which all the dots have the same color, and now try to pick out the clusters of populations that are evidenced in the tree. Not so easy, is it?

This is perhaps the biggest drawback of PC (and MDS and CA): there is (at present) no good objective way to identify the truly significant clusters of OTUs and distinguish them from clusters that are not supported. Instead, what is commonly done is to subjectively describe the clusters that you see in the plots, often aided by the affiliation of the OTUs (such as the continental affiliation of the populations in Figure 10.10), and then describe the results as if the plot provides strong evidence for the clusters—and I am not pointing fingers here because I am as guilty of this as everyone else! The bottom line is that it is generally useful to carry out both tree-based and PC (or MDS or CA)-based analyses of genetic data and compare the results, but at the same time you should beware of (or at least, be aware of the issues with) statements concerning the clusters that one "clearly" sees in the plots.

In addition to plotting PC components, they can also be visualized on a map, producing so-called **synthetic maps**. There are various ways to do this, but the basic idea is to calculate the PC1 values (for example) for your set of populations, plot the values on the map, and then use some method that involves interpolating between nearby values to provide a dense set of values across the entire geographic space of interest. Values can then be binned into different frequency classes and each class assigned a color; the result is something like you see on the cover of this book. Such synthetic maps are visually quite impressive (which is one reason why I chose this image for the cover!) and they provide a convenient visual overview of a huge amount of data. They are, however, not without significant concerns. First, a major concern has to do with how much of the patterns in the synthetic map reflect the data and how much reflect the interpolation. Suppose I have data from some populations from Egypt and some populations from South Africa; if I carry out PC, plot the PC1 values, and then do the interpolation, I'll inevitably get a nice smooth gradient in the change of PC1 all across Africa, even though I actually have data only from the

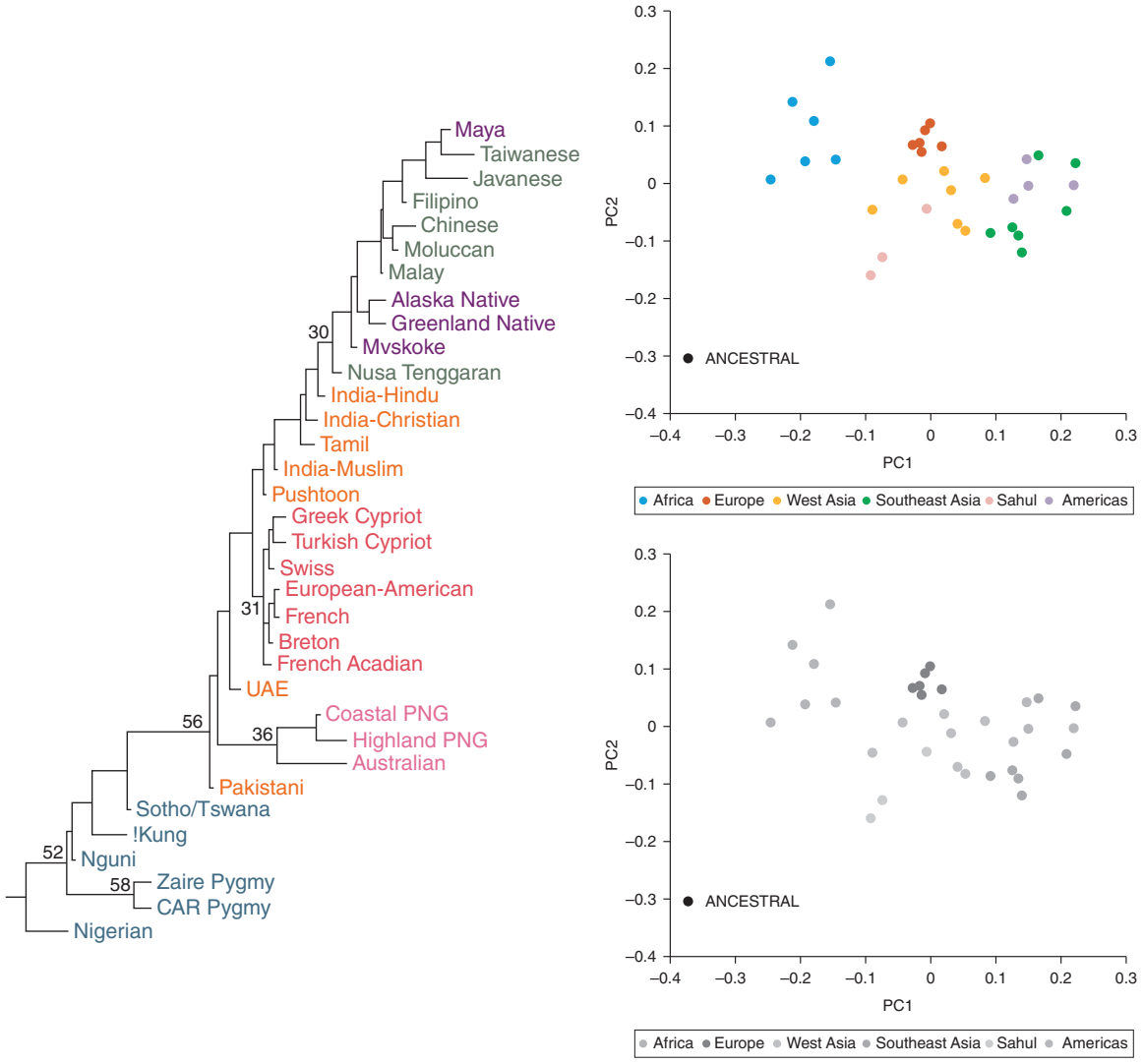


FIGURE 10.10

Example of a neighbor-joining (NJ) tree (left) and PCA plot (right) for the same data, consisting of 34 populations genotyped for eight Alu insertion polymorphism loci. The colors indicate the different geographic groupings of the populations. Bottom right is the same PCA plot, but without using different colors to indicate the geographic groupings. Reprinted with permission from Stoneking, M., et al., "Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa," *Genome Research* 11:1061, 1997.

extreme north and south and hence have no idea as to what the PC1 values actually are for the rest of Africa. For this reason, I tend to be extremely skeptical of such analyses when the actual sampling points are not indicated on the map.

Synthetic maps—like PC plots and trees—also lend themselves to overinterpretation. For example, Figure 10.11 is the map of PC1 in western Asia and Europe, based on analyses of classical markers by Luca Cavalli-Sforza and colleagues. There is a clear southwest to northeast cline in the values of PC1, which they argued reflects the demic expansion of farmers from western Asia to Europe. In other words, the map shows the spread of "farming genes" across Europe, and so farming spread not by cultural diffusion but by people migrating. However, it is difficult to make such associations between a synthetic map and prehistoric events, for several reasons. First, there is no temporal information in the map—the patterns in the map could indeed represent the spread of farmers across Europe beginning about 9000 years ago, or they could correspond to some completely different migration that occurred either earlier or later than the spread of farming. Second, there is no directionality inherent in the change in PC1—it could be a southeast to northwest migration, or vice versa. In fact, the change in PC1 evidenced in Figure 10.11 need not indicate any migration at all, as simulations have shown that similar

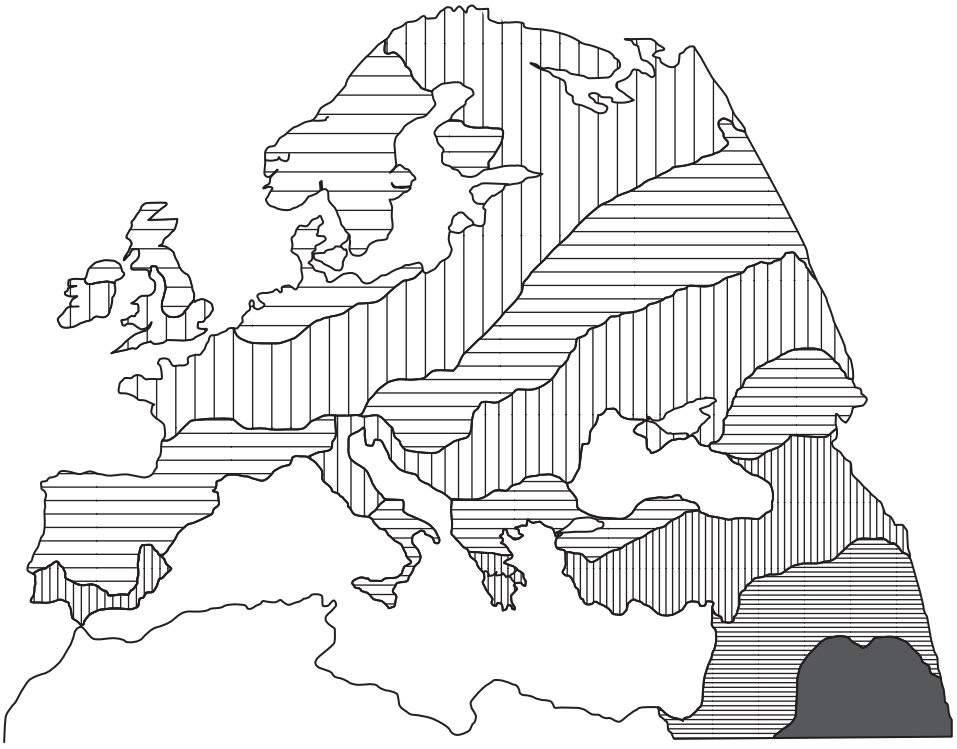


FIGURE 10.11

Synthetic map of PC1 values for Europe, based on classical genetic markers, illustrating a southeast to northwest gradient in PC1 values that may (or may not) reflect the spread of farmers across Europe. Modified with permission from Cavalli-Sforza, L.L., "Genes, peoples, and languages," *Proceedings of the National Academy of Sciences USA* 94:7719, 1997.

patterns can be produced under a model of isolation by distance (i.e., populations that are closer together geographically exchange genes more often than populations that are further apart geographically), without any long range migration of people (Novembre and Stephens 2008).

The important take-home message is that the analyses of populations described in this chapter (computing genetic diversity, genetic distances, AMOVA, using trees and plots to display the information in the data, etc.) are all useful descriptive analyses that can provide some ideas and hypotheses as to what might explain the observed patterns. However, if we want to go beyond mere description and storytelling, then we need methods to discriminate among potential competing explanations, and if we want to understand the important events that might have influenced current patterns of human genetic diversity, then we need ways of estimating demographic parameters of interest, like population divergence times, migration rates, and so forth. We will come to this soon—but first, we will consider analyses where the individual (sequence, haplotype, or multilocus genotype) is the unit of analysis.

CHAPTER 11

ANALYSIS OF GENETIC DATA FROM INDIVIDUALS

In the previous chapter, we covered the analysis of genetic data from populations. This is traditionally how such data have been analyzed, because for many years the only type of genetic data available consisted of allele frequencies determined for a sample of individuals from a population (after all, there is a reason why it's called population genetics!). But beginning in the late 1970s with the advent of methods for analyzing DNA variation, it became possible to analyze data at the level of the individual. For example, you might want to estimate genetic distances between pairs of individual DNA sequences or construct trees relating individual haplotypes or sequences. And more recently, with the advent of genome-wide single nucleotide polymorphism (SNP) typing and sequencing, it is possible to investigate the relationships among individuals based on their multilocus genotypes. Analyses where the individual is the unit of analysis, rather than the population, is the subject of this chapter.

GENETIC DISTANCES FOR DNA SEQUENCES

Just as we can calculate genetic distances between populations, we can also calculate a genetic distance between DNA sequences based on polymorphism data. As discussed in Chapter 7, the first DNA polymorphism data that became available were based on restriction fragment length polymorphisms (RFLPs). And soon afterward, genetic distances were designed to deal specifically with these kinds of data. However, because nobody carries out RFLP analysis anymore, there is little point in going over genetic distance measures that take into consideration the specific properties of RFLPs. Just be aware, should you read any of the "classic" papers (such as the early studies of human mtDNA variation) that are based on RFLP data, that specific genetic distance measures are required for such data—you can't use just any old genetic distance measure (such as F_{ST}).

Nowadays, DNA sequences are all the rage, so it is worthwhile discussing genetic distance measures that are designed specifically for DNA sequences. Suppose we have DNA sequences from the same region of the genome from two individuals and we want to estimate how different they are. Intuitively, you might think, well, just count the number of nucleotide positions where they differ and divide by the total number of positions in the sequence; that will give you the proportion of nucleotide differences between the two sequences. However, a potential problem with this approach becomes evident if you consider what you will get if you write down two DNA sequences as a random series of nucleotides—they won't differ by 100%, as you might expect for two completely unrelated sequences, but rather by about 75% (on average), because there is a one in four chance that two unrelated sequences will have the same nucleotide at the same position. Moreover, with mutations occurring at random, not all mutations that have occurred during the past will be seen as a difference between two sequences. For example, as shown in Figure 11.1, there can be multiple substitutions at the same site in one lineage, mutations from an ancestral nucleotide to a derived nucleotide and then back to the ancestral nucleotide (**back mutations**), and independent mutations in two different lineages at the same site to the same nucleotide (**parallel mutations**). The overall result: the observed number of nucleotide differences will underestimate the actual number of nucleotide substitutions that have occurred between two sequences. So how can we use what we can observe (the number of nucleotide differences between two sequences) to figure out what we would

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

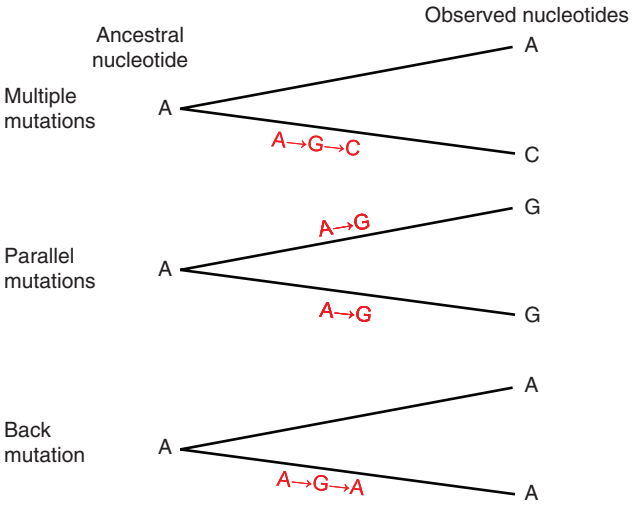


FIGURE 11.1

Examples of multiple, parallel, and back mutations. Note that in each case two mutations have actually occurred but fewer mutations (either none or one) are observed.

actually like to know (the number of mutations, or nucleotide substitutions, that have occurred since our two sequences last shared a common ancestor)?

Let's start with a simple model. Suppose there is a single rate, α, for all substitutions that can occur in a DNA sequence. We can then write out the substitution matrix as follows:

	A	G	C	T
A	—	α	α	α
G	α	—	α	α
C	α	α	—	α
T	α	α	α	—

In this sort of matrix, the row gives the original nucleotide, and the column gives the new nucleotide after the mutation. Now, let's focus our attention on a particular nucleotide position in our sequence in a particular generation, t. What is the probability of finding an A at this position? This could happen in two ways. First, there was also an A at this position in the previous generation ($t - 1$), and there was no mutation, so it is still an A in the present generation. If we denote the probability of an A at this position in generation t as P_t, and we note that the overall probability of A mutating to some other nucleotide is 3α (because A can mutate to G, C, or T, each with probability α, so the total chance of a mutation is $\alpha + \alpha + \alpha$), then the chance of an A staying an A is the probability that there was an A at this position in the previous generation and it did not mutate, which is $(1 - 3\alpha)P_{t-1}$. The second way we can have an A in generation t is if there was some other nucleotide at this position (G, C, or T) in generation $t - 1$ and there was then a mutation to A. The chance of having a non-A nucleotide at this position is $1 - P_{t-1}$, and the chance of a mutation specifically to A is α, so the overall probability of a non-A nucleotide mutating to an A is $\alpha\ (1 - P_{t-1})$. Let F be the fraction of nucleotide positions that are different between two sequences, which is what we can actually observe. And, let k be the actual number of substitutions that have occurred per position, which is what we want to know. Then with this model (and a lot of math that we'll skip over because there is nothing of interest to be learned by doing the math), it can be shown that:

$$k = -3/4\ (\ln(1 - 4F/3)$$

where "ln" is the standard abbreviation for the natural logarithm, which is logarithm to the base e (as opposed to logarithm to the base 10, which was mentioned in Chapter 5). e is an irrational number, like π, meaning that the numbers after the decimal never end or form a pattern, and is approximately equal to 2.71828. First studied in detail by the Swiss mathematician Leonhard Euler in the 1720s (the fact that e is also the first letter of his last name is probably not a coincidence!), e pops up in all sorts of places in science and nature (which is why it is the "natural" mathematics constant), including processes involving nonlinear rates of increase or decrease (population growth, interest rates, etc.), the shapes of arches or hanging cables, and some types of probability distributions. For example, if you have a 1 in n chance of winning the lottery, and you play the lottery exactly n times, then (if n is large enough, more than 20 or so), the probability that you do not win is about $1/e$, or about 37%.

Anyway, the above equation is known as the "Jukes–Cantor" equation, as it was developed by Tom Jukes and Charles Cantor (Jukes and Cantor 1969). Incidentally, together with Jack King, Jukes independently came up with the idea of neutral theory that was attributed to Motoo Kimura back in Chapter 6. Although Kimura's paper came out while they were still writing theirs (Kimura 1968), they still submitted their paper to the prestigious journal *Science*, where it was promptly rejected, one reviewer stating that the results were too obvious and trivial to be worth publishing, the other reviewer stating that it was all wrong (Crow 2000). Fortunately, they appealed and managed to get their paper published (King and Jukes 1969)—yet another example of a seminal paper that almost didn't see the light of day.

Note that the rate parameter, α, does not appear in the equation (α affects both k and F equally, so it cancels out), which is a relief because we usually not only don't know what α is, we often don't even have a good way to estimate it. Figure 11.2 is a graph of F (the observed fraction of nucleotide positions that

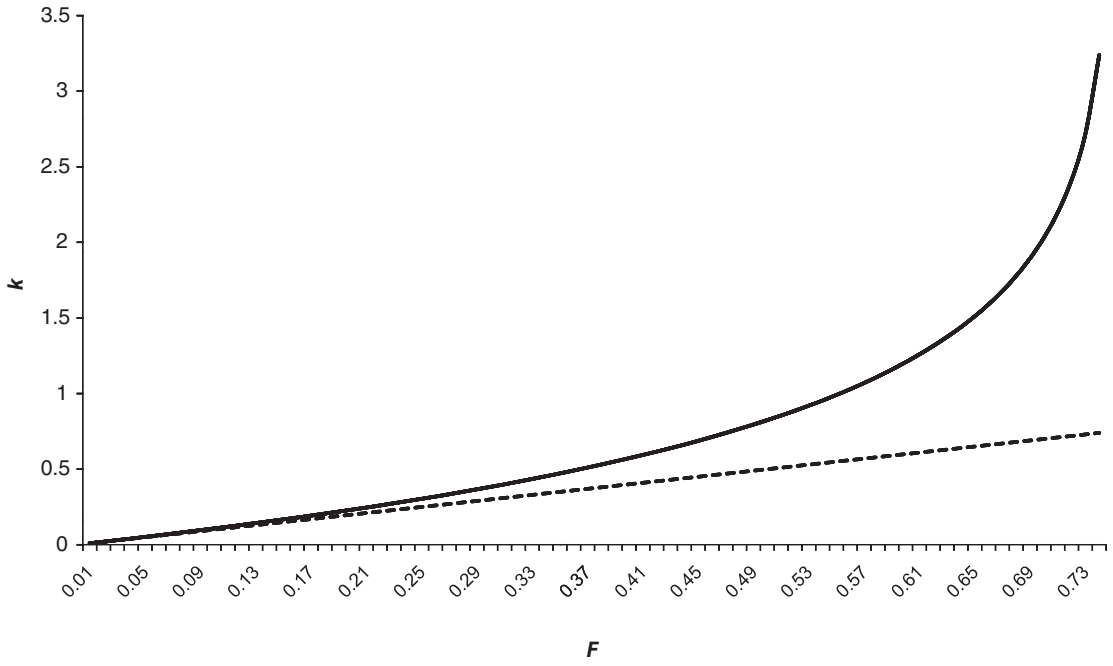


FIGURE 11.2

Graph of F (the observed fraction of sites that differ between two sequences) versus k (the estimated number of nucleotide substitutions per site) for the Jukes–Cantor model. Dashed line, with no correction; solid line, with Jukes–Cantor correction.

differ between two sequences) versus k (the actual number of nucleotide substitutions per site); F is also known as the "corrected" distance, or (in this case) the Jukes–Cantor correction, as it can be thought of as the observed distance corrected for multiple substitutions at the same site. As shown in the graph, for values of F greater than about 0.25, the observed differences between two sequences is an underestimate of the actual number of nucleotide substitutions that have occurred, with the discrepancy becoming bigger with bigger values of F. For example, when $F = 0.67$ (corresponding to an average of 2/3 of a mutation per site), the actual number of mutations that have occurred is about 1.5. So, for distantly related sequences the correction is quite important, but for closely related sequences ($F < 0.25$ or so), there is no need to worry about multiple substitutions, parallel mutations, or back mutations, which hopefully makes sense as over short evolutionary time periods there should be very few (if any) such events.

The model we developed assumes that all mutations occur with the same frequency—how realistic is this assumption? The answer is: not very realistic at all. It turns out that two of the nucleotides found in DNA, A and G, are classified as **purines** based on their chemical composition, while the other two nucleotides, C and T, are classified as **pyrimidines**. If all mutations occur at the same frequency, then we would expect **transversions** (mutations involving an exchange of a purine for a pyrimidine or vice versa) should occur twice as often as **transitions** (mutations involving an exchange of one purine for another or one pyrimidine for another). For example, an A can mutate to a G (transition), C (transversion), or T (transversion). In fact, transitions occur more frequently than transversions: in nuclear DNA, there is a roughly 2:1 ratio of transitions to transversions, while in mtDNA the ratio approaches 10:1. This excess of transitions probably reflects the fact that exchanging one purine for another, or one pyrimidine for another, does not have as big an impact on the DNA structure as exchanging a pyrimidine for a purine (or vice versa). That is, transitions involve chemically similar nucleotide bases and hence probably are not as readily recognized as errors by the proofreading mechanism of the DNA polymerase.

Thus, it would be more realistic to have a model that allows for unequal rates of substitution for transitions versus transversions. Fortunately, that is easy enough to do; let α be the mutation rate for transitions and β be the mutation rate for transversions, then the substitution matrix becomes:

	A	G	C	T
A	—	α	β	β
G	α	—	β	β
C	β	β	—	α
T	β	β	α	—

If we designate the fraction of positions that differ by a transition as P, and the fraction of positions that differ by a transversion as Q (so P + Q = F), then by following the same reasoning as before (i.e., an A at a position in the current generation can either be an A that did not mutate in the previous generation, or some other nucleotide that mutated to an A) and with a lot of math, we can get a really complicated equation, too complicated to write out here. The important point is that with this equation, we can estimate what *k* is, based on what we can actually observe: namely, the number of transition and transversion differences per site (P and Q, respectively). This model is known as the Kimura 2-parameter model of DNA sequence evolution, as it was first developed by Motoo Kimura (of neutral theory fame, mentioned back in Chapter 6) in 1980 (Kimura 1980). An important point to remember is that if there is no bias in the rate of transitions versus transversions, then the Kimura 2-parameter equation reduces to the Jukes–Cantor equation. Figure 11.3 shows the relationship between *F* and *k* for different ratios of transitions to transversions (keep in mind that *F* = P + Q). As with the Jukes–Cantor equation, it is only when *F* is about 0.25 or bigger that there is any difference between the resulting *k* values and the observed fraction of differences per position. However, as the transition bias gets bigger and bigger, the difference between *F* and *k* also gets bigger and bigger, meaning that our observed *F* seriously underestimates the actual number of mutations that have occurred. This is because more and more of the mutations are transitions, and so there is a greater tendency for a base to mutate back and forth between the two purines or between the two pyrimidines, for example:

$$A \rightarrow G \rightarrow A \rightarrow G \ldots \text{ or } C \rightarrow T \rightarrow C \rightarrow T \ldots$$

In such cases, we observe only (at most) a single transitional difference, when in fact there have been many transitional substitutions, and hence what we observe greatly underestimates the actual number of substitutions that have occurred.

We can also introduce additional parameters to allow for different mutation rates for A ↔ G versus C ↔ T transitions or for the various kinds of transversions, if we have reason to think that such differences might exist. The resulting equations for k then get very messy very quickly, so we won't bother going through them.

However, another important assumption of these models is that the rate of mutation is the same for all positions; how realistic is this assumption? If we look at mutation rate estimates at different sites across the mtDNA genome, there seems to be substantial variation (Figure 11.4). But is this really more variation than would be expected by chance? After all, since mutation is a random process, we would expect some variation in mutation rates from site to site. One way to look at this question is to examine the inferred number of mutations at each position (which we can get from phylogenetic or network analyses, discussed later in this chapter) and ask whether the distribution of the number of mutations per position differs significantly from what would be expected if the mutation rate was the same for each site. Table 11.1 gives some data for 20 human mtDNA control region sequences consisting of 1116 positions, where it was estimated that

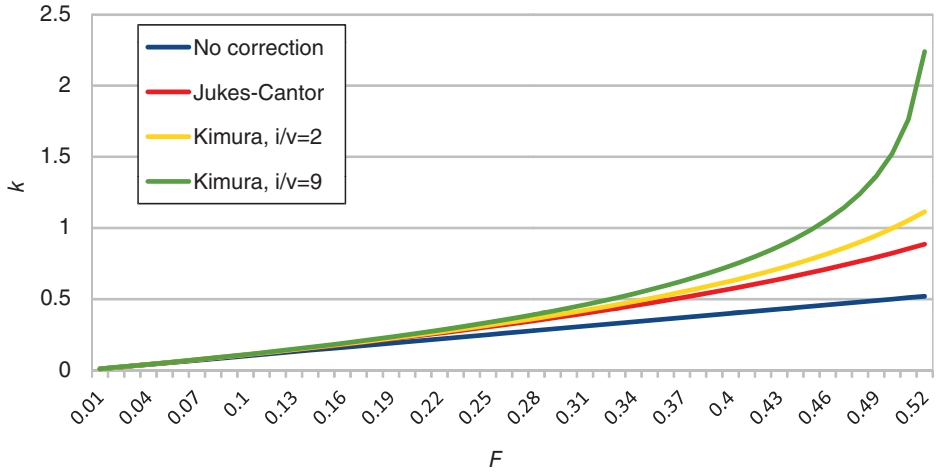


FIGURE 11.3

Graph of *F* versus *k* for no correction, Jukes–Cantor correction, Kimura correction with ratio of transitions to transversions (i/v) = 2, and Kimura correction with i/v = 9. Note that for small values of *F* (corresponding to closely related sequences) the difference between *F* and *k* is negligible, but as *F* gets larger and larger, *F* underestimates *k* by a larger and larger factor, especially if there is a bias for transitions versus transversions.

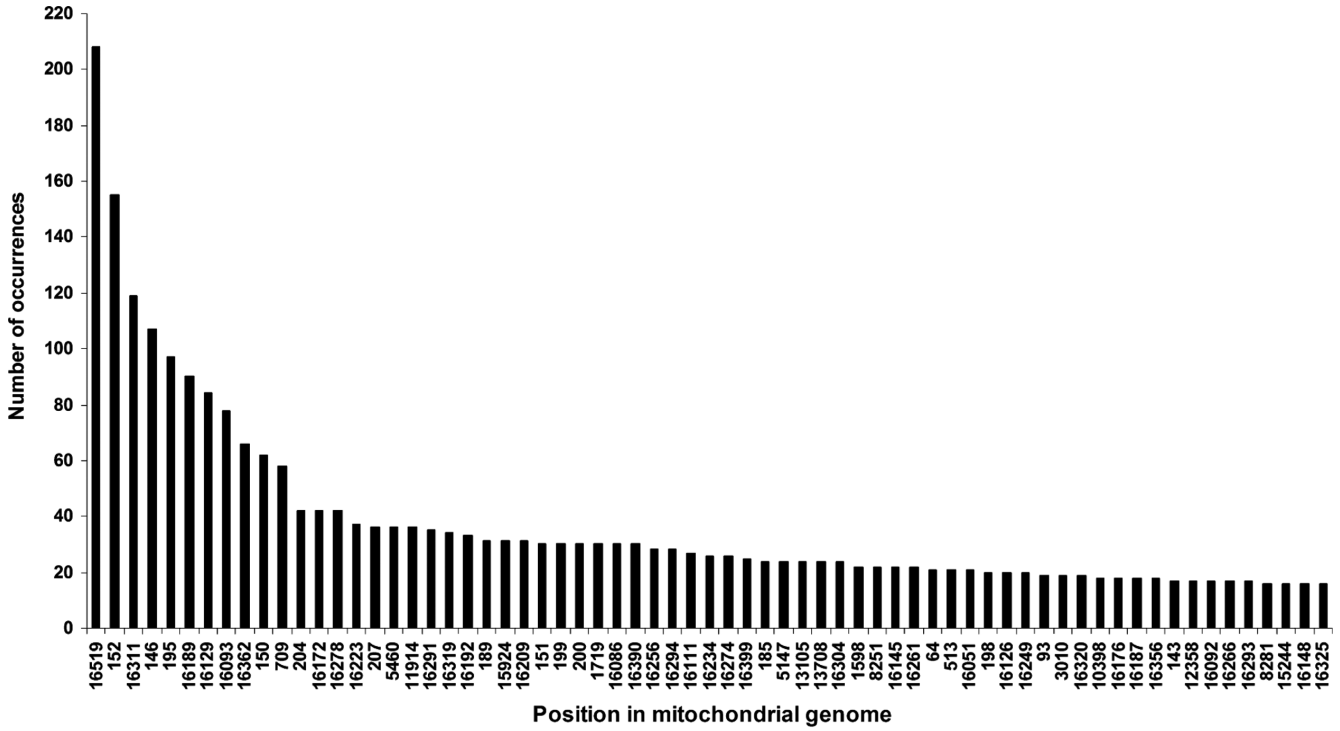


FIGURE 11.4

Bar chart of the relative mutation rates for the most rapidly evolving sites across the human mtDNA genome. The X axis gives the nucleotide position for those sites with 25 or more observed mutations (the Y axis gives the number of mutations inferred at that position), based on a phylogenetic analysis of 2196 complete mtDNA genome sequences. Reprinted with permission from Soares, P., et al., "Correcting for purifying selection: an improved human mitochondrial molecular clock," *American Journal of Human Genetics* 84:740, 2009.

1028 positions had not mutated at all, 58 positions had mutated once, 21 positions had mutated twice, and 9 positions had mutated three or more times. Our expectation is that if the mutation rate is the same for each site, then the distribution of the number of mutations per position should follow a Poisson distribution (the Poisson distribution was mentioned previously back in Chapter 3, in the section on the effect of differential fertility on the effective population size). If we compare the observed number of mutations per position to the expected number assuming a Poisson distribution (Table 11.1), they are in fact significantly different. But if we instead assume that the mutation rate is not the same at each position, then it turns out that we can get a good fit between the observed and expected number of mutations per position (Table 11.1). It's as if we were making chocolate chip cookies by first preparing the batter and then adding chocolate chips to the batter and stirring everything together. If we are a good baker and stir the batter quite thoroughly, then the number of chocolate chips per cookie should follow a Poisson distribution. But if we are a lazy baker and don't stir the batter very well, then we end up with both more cookies with lots of chocolate chips and more cookies with hardly any chocolate chips than we would get if we had stirred the batter thoroughly. And that is what we observe in Table 11.1: more sites with no mutations, and more sites with many mutations, than would be expected under a Poisson distribution. So, evolution is a lazy baker!

Why there should be such variation among mutation rates is, for the most part, a mystery. However, there is one well-known example of mutational

TABLE 11.1 ■ Observed number of substitutions per site, based on parsimony analysis of 20 human mtDNA control region sequences, and expected numbers for a model either without rate variation (Poisson) or with rate variation[a]

Number of substitutions per site	Number of sites		
	Observed	Poisson	With rate variation
0	1028	987.0	1028.0
1	58	121.2	59.3
2	21	7.4	17.4
≥3	9	0.4	11.3

[a]From Tamura and Nei (1993).

hotspots where we do understand the mechanism, and that is the case of CpG dinucleotides, which are C followed by G in the same DNA strand (the "p" stands for the phosphate bond that links two adjacent nucleotides on the same DNA strand). This is not to be confused with the usual base-pairing between C on one strand of the DNA double helix and G on the other strand of the double helix. Wherever a C is followed by a G on the same DNA strand, there is a higher chance of the C mutating to a T than there is when the C is followed by some other nucleotide (C, A, or T). This is because the C in CpG sites frequently has a methyl group (a carbon atom bonded to three hydrogen atoms) added to it; such methylation is an important part of cell growth and development, as it influences patterns of gene expression and hence contributes to the differentiation and maintenance of distinct tissues within the body. It turns out that methylated cytosine (C) nucleotides can undergo deamination (loss of an amine group, which is a nitrogen atom bonded to three hydrogen atoms) to form thymine (T), so there is a high rate of C → T transitions at methylated CpG sites. This results in the loss of CpG sites over evolutionary time, and in fact the frequency of CpG sites in the human genome is only about 20% of what would be expected based on the overall frequency of C and G nucleotides.

So, it would be useful to take into account such site variation in mutation rates in our estimate of the number of substitutions between two sequences. Fortunately, methods have been worked out to do this. It turns out that the observed rate variation among sites can be well modeled by a gamma distribution, which is a mathematical distribution whose shape is determined by the **gamma parameter**, *a* (also known as the shape parameter). Figure 11.5 shows the distribution of relative evolutionary rates associated with different values of *a*. Notice that as *a* gets bigger and bigger, the distribution tends to be more and more centered around a single evolutionary rate—when *a* is infinity, then there is no rate variation and all sites have the same evolutionary rate. As *a* gets smaller and smaller—in particular, when *a* is less than one—then there are more sites with low evolutionary rates, and more sites with high evolutionary rates, compared to larger values of *a*.

We can then incorporate rate variation among sites into our estimate of *k* from the Kimura 2-parameter model, which leads to an even more complicated expression (which again there is no point in showing). The main thing to remember is that we can use this approach to estimate *k*, based on our observed transition and transversion differences (P and Q), as well as an estimate of *a*. For human mtDNA control region sequences, *a* has been estimated to be around 0.11 (Kocher and Wilson 1991; Tamura and Nei 1993), and Figure 11.6 shows the relationship between *k* and

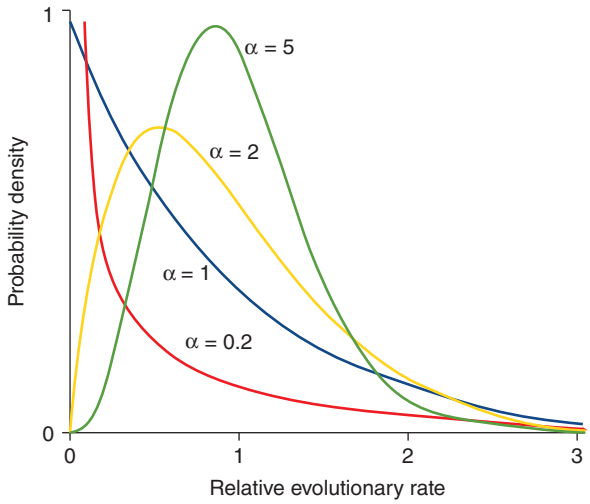


FIGURE 11.5

The effect of different values of the *α* parameter on the distribution of evolutionary rates across sites. As *α* becomes smaller and smaller, the distribution becomes more and more skewed, with many sites with a low evolutionary rate and a few sites with a high evolutionary rate.

F (again, *F* = P + Q) for various values of *a*. As we saw previously, for small values of *F* the correction for unequal rate variation is not very important, but for large values of *F* it can indeed be very important. In principle, one can incorporate rate variation among sites into any model of nucleotide substitution by including the *a* parameter. In practice, most people who include rate variation in their model will use Tamura–Nei gamma distances (Tamura and Nei 1993), because these also allow for unequal frequencies of the four nucleotides (the Jukes–Cantor and Kimura distances described previously do assume equal frequencies of all four nucleotides). However, as with the Kimura equations we won't bother showing the formula for Tamura–Nei gamma distances as it is an extremely complex expression and there is nothing to be gained by showing it. There exist many computer programs that implement all of these various distance measures (and more), so they are easy enough to calculate without worrying about the formulas.

In the end, how important are these various extensions (transition bias, unequal mutation rates across sites, etc.) to how we estimate the number of substitutions between two sequences? It all depends on the properties of the sequences under examination. If the sequences are relatively closely related (say, *F* is less than 0.25) and there isn't much variation in the mutation rate across sites, then you'd be better off using the observed number of differences divided by the total number of sites in common in

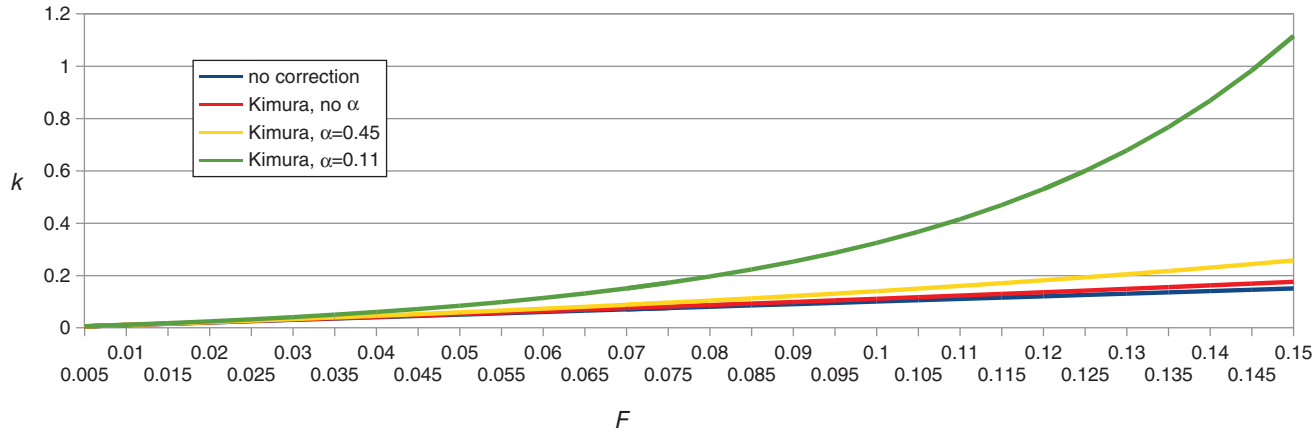


FIGURE 11.6

Graph of F versus K for the Kimura model (with i/v = 9) with no correction for rate variation, and for rate variation with $\alpha = 0.11$ and $\alpha = 0.45$. Note that the scale on the X axis (for F) is different from that in Figure 11.3, further emphasizing the large impact of rate variation on estimates of k from F.

the two sequences (this sometimes goes by the name **p-distance**). Not only is it simpler to calculate, it also has the lowest variance of any estimate of sequence difference, so you'll get a more precise answer. But if your sequences evolve by a more complicated model, then you need to take the complications into account or you run the risk of getting a very wrong answer. For example, as noted before, the human mtDNA control region exhibits a very strong transition bias and high variability in substitution rates across the region. Suppose we want to date the age of the human mtDNA ancestor from control region sequences, using a molecular clock approach (why we might want to do this will be discussed in Chapter 14). We've built a tree for our human mtDNA sequences (using either methods discussed in the previous chapter or in the next section) and gotten an estimate of 2.87% (i.e., almost three mutations in every 100 base-pairs of the mtDNA control region) for the amount of sequence divergence that has accumulated since the common human mtDNA ancestor (Vigilant et al. 1991). In order to convert this estimate into time, we need to know the substitution rate for mtDNA, so we compare our human control region sequences to a chimpanzee control region sequence. We observe an average of 15.1% sequence divergence (i.e., about 15 differences in every 100 base-pairs between the human and chimpanzee control region sequences). If we assume that humans and chimpanzees diverged 5 million years ago, then the rate of mtDNA control region sequence divergence is 15.1/5 = 3.02% divergence per million years, and so the age of our human mtDNA ancestor is estimated to be (2.87/3.02) × 1 million years, which is about 950,000 years. If we instead use the Jukes–Cantor distance, the divergence between humans and chimpanzees becomes 16.8%, and the resulting estimate of the age of the human mtDNA ancestor is slightly more recent, about 850,000 years ago. The Kimura distance (without any gamma correction) doesn't change things much, as the divergence between humans and chimpanzees then becomes 17.3%, and the age of the human mtDNA ancestor is then estimated to be about 830,000 years ago. But if we use the Tamura–Nei gamma distances, the human–chimpanzee divergence is much bigger, about 75.2%, and the age of the human mtDNA ancestor is much more recent, about 190,000 years ago. Clearly, the rate heterogeneity in the mtDNA control region has a big impact on our estimates of sequence divergence (and remember, if this were not the case, then there wouldn't be much difference between estimates of sequence divergence that allow for rate heterogeneity versus estimates of sequence divergence that assume equal rates of substitution at all sites). The take-home message: if you want the right answer, you've got to use the right model.

TREES FOR DNA SEQUENCES

In the previous chapter, we considered how to build trees from genetic distance matrices, such as unweighted-pair-group-method-of-averaging and neighbor-joining trees. In such trees, the operational taxonomic units (OTUs) were populations. We could use these same methods to construct trees for individual DNA sequences, where the genetic distances are the estimated number of substitutions between each pair of sequences, obtained by using one of the methods described in the previous section. In such trees, the OTUs are the individual DNA sequences, and indeed this is a common way of constructing trees based on

	Position			
OTU	1	2	3	4
w	G	C	T	A
x	G	T	C	A
y	T	T	T	G
z	T	T	C	G

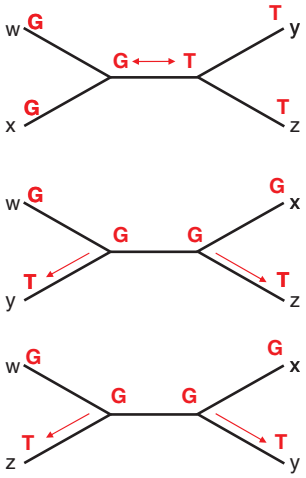


FIGURE 11.7

Example of maximum parsimony analysis. The box shows the nucleotides at four positions in sequences from four operational taxonomic units (OTUs); on the right are the three possible unrooted trees for these four OTUs. The nucleotide positions in each tree correspond to the first position in the sequences (red color). See text for further details and explanation.

DNA sequences. However, whenever we construct distance matrices, we inevitably lose information, as a single number (the genetic distance) does not capture all the information in the data. It turns out that there are other methods for constructing trees from DNA sequences that are based on **character state** information, that is, where the data used to construct the tree consist of the particular nucleotide present at each position in each sequence, and these tend to be more informative.

How do such methods work? Let's start by considering the sequences in Figure 11.7. There are four sequences (W, X, Y, and Z—these are our OTUs) each with four positions. For these four OTUs there are three possible trees (regardless of how long the sequences are—for four OTUs there are always only three trees) that depict how they are related, also shown in Figure 11.7. These three possibilities group W with X and Y with Z, W with Y and X with Z, and W with Z and X with Y. Note that these are all of the possible trees for four OTUs (i.e., the possible unrooted trees—how one goes about identifying the root, or ancestor of everyone, in the tree is discussed in the "Rooting trees" section). Now, consider the position shown in red boldface in the sequences: W and X both have a G at this position, while Y and Z both have a T. If we assume that Tree #1 is the true history of these four OTUs, then we can explain the data for this position with one mutation: the common ancestor of W and X would have a G, the common ancestor of Y and Z would have a T, and there would have been a G ↔ T change on the branch linking these two ancestors.

Now let's look at Tree #2, which links W with Y and X with Z. Here it turns out that if this tree represents the true evolutionary history of these four OTUs, then there must have been (at least) two mutations at this position in order to get the observed sequences. There are various ways that the two mutations could have occurred, and the figure shows one of them: the common ancestor of W and Y had a G at this position and there was a mutation from G → T on the lineage leading to Y, and the common ancestor of X and Z also had a G at this position, and there was a mutation from G → T on the lineage leading to Z (this would be an example of a parallel mutation). Although we don't know for sure what happened at this position, we do know that if Tree #2 is the true history of these sequences, then there must have been at least two mutations.

And what about Tree #3, linking W with Z and X with Y? Hopefully, it is clear to you that Tree #3 is exactly like Tree #2 in terms of the character (nucleotide) states at this position, and so like Tree #2 this tree requires at least two mutations. Therefore, by the principle of **maximum parsimony**, which basically states that the potential evolutionary history that requires the least amount of change (i.e., is most parsimonious) is likely to be the true history, we would conclude that Tree #1 is most likely to represent the true history of these sequences as it requires the fewest number of mutations. Maximum parsimony sounds a bit like an oxymoron—after all, what does it mean to "maximize" the least amount of change (an oxymoron, for those of you who don't know, is a term that puts together two or more words that seem inherently contradictory, like "jumbo shrimp" or "military intelligence"). It also may seem somewhat arbitrary, as how do we know whether evolution really does proceed by the pathway that requires the least amount of change? This point has been discussed a great deal, especially by those who study the evolution of morphological or skeletal traits (for which maximum parsimony was first used). However, in molecular evolution, it has been shown that if mutations are rare (which, for the most part, they are), then maximum parsimony is in fact expected to get you the right answer.

There are some other features of maximum parsimony that should be pointed out. First, not all

positions are informative when it comes to figuring out which tree has the shortest length. For example, look at the second position in the sequences in Figure 11.7. Here, the first sequence has a C, and the other three sequences have a T. If we now fit this position to the three possible trees for these four OTUs, we find that this position always requires one mutation, regardless of the tree, if we assume that there was a change from T → C on the branch leading directly to OTU W. Therefore, such positions are not informative in a maximum parsimony analysis; the sites that are informative (called, aptly enough, **phylogenetically informative sites**) are polymorphic sites with the minor (less frequent) allele present in at least two OTUs. Hopefully, it is also clear that sites that are not polymorphic (i.e., where everyone has the same allele or base) are also not phylogenetically informative!

Another important point concerning parsimony is that different positions may support (i.e., have the shortest length for) different trees. Consider the third position in the sequences in Figure 11.7, in which OTUs W and Y have a T and OTUs X and Z have a C. For this position, the second tree in Figure 11.7 would require only one mutation, while the first and third trees would require (at least) two mutations. So, if we have only two phylogenetically informative sites in our sequences, one like the first position that favors Tree #1, and one like the third position that favors Tree #2, what do we do? Clearly, they can't both be right, so one (or both) of these positions must have undergone two (or more) mutations. With real data, conflicts between sites always occur—you never (well, hardly ever) have real data for which the best tree requires only one change at every polymorphic position; back mutations and parallel changes at the same site in different lineages are not exactly common, but they are also not exceptional. And there are basically two ways of handling such conflicts. One is the "egalitarian" (i.e., majority rule) approach, which takes the view that all sites are equally informative, so you simply count the total number of mutations required by each tree and choose the one(s) with the fewest number of mutations. In our artificial example in Figure 11.7, if we look at all four positions in the sequence and focus only on the phylogenetically informative sites (sites 1, 3, and 4), then Tree #1 requires four changes at these three sites, while Tree #2 requires five changes and Tree #3 requires six changes at these three sites. Therefore, Tree #1 "wins" and is our best estimate of the true evolutionary history of these sequences. And if we had only the first three positions in the sequence? Then by the "egalitarian approach" we would have two equally likely possibilities, Tree #1 and Tree #2, both requiring three changes at the two phylogenetically informative positions in the sequences, so we would conclude that these are equally parsimonious trees and hence we cannot distinguish between them with the data at hand. In reality, it is quite common to have numerous equally parsimonious trees for a given data set, so what one then does is focus on the patterns that are present (or absent) in all of the equally parsimonious trees—in the example in Figure 11.7, we don't know whether W is most closely related to X or to Y, but we do know that it is *not* most closely related to Z.

The other way to handle conflicts among sites is the "elitist" approach, in which one takes the view that some sites are better (i.e., more informative) than others when it comes to trying to find the best tree. For example, as mentioned previously, it is well-known that transitions occur much more frequently than transversions, especially in mtDNA, so we might take the view that we should minimize not only the total number of mutations but also the total number of transversions required by a tree. In the example in Figure 11.7, if we look at only the first and third positions in the sequences, then as we saw previously we have two equally parsimonious trees that require three mutations, Tree #1 and Tree #2. However, Tree #1 requires one transversion (at the first position) and two transitions (at the third position), while Tree #2 requires two transversions and one transition. Hence, if we take the view that transversions are likely to occur less often than transitions, we should favor Tree #1 as our best estimate of the evolutionary history of these sequences. This sort of "elitist" approach can be quite useful—as long as you have good reason to give more weight to certain kinds of mutations.

Figuring out which tree is the best according to maximum parsimony is straightforward enough for four OTUs, but what if we have more? It turns out that as we increase the number of OTUs, the number of possible trees increases enormously: for 10 OTUs there are more than 34 million (rooted) trees, and for 50 OTUs there are about 2.75×10^{76} different trees, which is approaching the number of atoms (about 10^{80}) in the known universe! Basically, once you get beyond about 15–16 OTUs or so, it is no longer computationally feasible to find the shortest tree by evaluating the lengths of all possible trees. So what do you do then? It turns out that there are **heuristic algorithms** that will build a tree using the maximum parsimony principle from the ground up by successively linking together OTUs while trying to minimize the total length of the tree. These algorithms make it possible to construct maximum parsimony trees for even hundreds of OTUs (although, depending on the complexity of the data, computation time also becomes a consideration). With these algorithms, one must keep in mind that there is no guarantee that the resulting tree really is the one requiring the fewest number of mutations—with simulated data the

heuristic algorithms generally get you pretty close to the right answer, but there is always the worry that real data differ from simulated data in some important but unknown feature. But even with these caveats, maximum parsimony remains an important, useful, and widely used method for producing trees from molecular data.

ROOTING TREES

The trees in Figure 11.7 may look kind of peculiar as they are not drawn as typical trees, with a root (or common ancestor) for all of the sequences. For each of the three unrooted trees, there are five possible rooted trees, as the common ancestor could be placed on any of the five branches in the unrooted tree; the 15 possible rooted trees for our four sequences are shown in Figure 11.8. So which of these 15 rooted trees most likely explains our sequences? It turns out that with the information we actually have—namely, just these four sequences—we cannot determine where the root should go. Even though unrooted Tree #1 is the preferred tree by maximum parsimony, we can place the root on any of the five branches in this tree, and all of these five rooted trees explain the data equally well as they all require the same number of mutations. If we want to put a root on our tree, then we need more than just the four sequences: we either need more information—for example, an **outgroup** sequence—or we must be willing to make additional assumptions—for example, assume a molecular clock.

An outgroup is an OTU for which you have independent knowledge that it is not more closely related to any of the OTUs in your analysis (the **ingroup** OTUs) than these ingroup OTUs are to one another. In other words, all of the ingroup OTUs in your analysis should share a more recent common ancestor with one another than any do with the outgroup OTU. If that is indeed the case, then by definition the root of the tree is where the outgroup OTU attaches to the unrooted tree. But if your supposed outgroup OTU is actually an ingroup OTU, then you are quite likely to come to the wrong conclusion as to where the root of the tree belongs, so it is very important that you have unequivocal evidence that your outgroup OTU really is an outgroup.

In our example in Figures 11.7 and 11.8, if the outgroup is attached to the branch leading directly to sequence W, then the corresponding rooted tree would be the one outlined in red. For most applications in molecular anthropology, where the OTUs are molecular data from various human populations, the preferred outgroup is the chimpanzee, because (as we shall see in Chapter 13) chimpanzees are our nearest living relatives. In principle, anything that is outside the range of human variation could be used as an outgroup (even a fruit fly or yeast), but in practice the more distantly related the outgroup, the greater the potential for incorrect rooting because of parallel

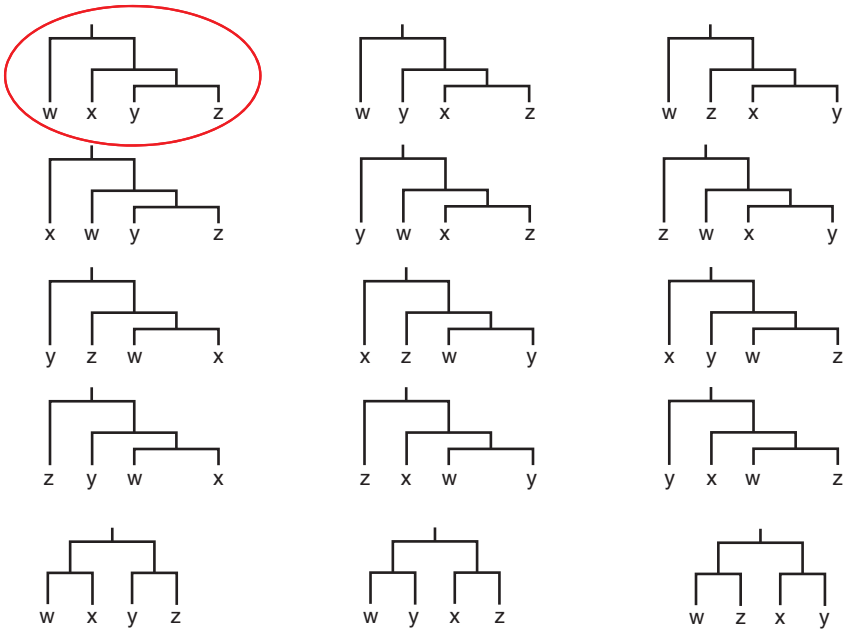


FIGURE 11.8

The 15 possible rooted trees for the unrooted trees shown in Figure 11.7. The tree circled in red corresponds to a rooting of Tree #1 shown in Figure 11.7, with the root attached to the branch leading to operational taxonomic unit w.

or back mutations—so, best to use the most closely related outgroup you can find.

What if you don't have the data you need from a good outgroup? For some types of polymorphisms, it is possible to deduce the ancestral state from the evolutionary properties of the polymorphisms. For example, Alu insertion polymorphisms (introduced in Chapter 6) are polymorphisms for the presence or absence of an Alu element at a specific chromosomal location. Because the direction of change is always the insertion of an Alu element into a new location, this means that the ancestral state for such polymorphisms is the absence of the Alu element. Therefore, a "hypothetical ancestral population" in which the frequency of each Alu element is zero can be included in the analysis (e.g., Figure 10.10), and the point where this hypothetical ancestral population attaches to the tree of ingroup OTUs is, by definition, the root of the tree.

When all else fails—that is, you don't have data from an outgroup, and you cannot otherwise infer ancestral states for your polymorphisms—then you can always fall back on midpoint rooting, as discussed in the previous chapter. Just remember that midpoint rooting does assume a molecular clock—that is, a constant rate of change in the data that were used to construct the tree—so you should then be sure to test that your data really do conform to the assumptions of a molecular clock (how to do this is described in Chapter 12).

ASSESSING THE CONFIDENCE OF A TREE

Once we have our best tree (or, more likely, our set of best trees), one question that then frequently arises is how much confidence we should have in the groupings in the tree. After all, there's not much point in trying to explain or emphasize a particular group of OTUs in a tree if it turns out that there is only weak support for that group, and other branching patterns are nearly as equally likely. One method that is commonly used to assess how strongly the data support the groups in a tree is **bootstrap analysis** (discussed already in the context of allele frequency data in the previous chapter). Applying bootstrap analysis to sequence data is quite straightforward: if your tree is based on N sequences (or OTUs), each consisting of X nucleotide positions that you've sequenced, then the data can be represented as a matrix with N rows (with each row corresponding to an individual sequence) and X columns (with each column corresponding to an individual nucleotide position)—for example, this is how the sequences are represented in Figure 11.7. To carry out the bootstrap analysis, you create a bootstrap data matrix by copying a column (nucleotide position) randomly from the observed data matrix, and doing this X times. This is known as sampling with replacement, and the idea is that you end up with a bootstrap data matrix with the same number of sequences and nucleotide positions as the original data matrix. This bootstrap data matrix is similar but not identical to the original data matrix, as some nucleotide positions are included more than once and others are not included at all. You then construct a tree for the bootstrap data matrix and see how many of the groups in the original tree are represented in the bootstrap tree. You then repeat this procedure a large number of times—usually, 1000 to 10,000 times—to generate a bootstrap sample of trees. If a particular grouping in the tree based on the original data is supported by a large number of nucleotide positions, then that grouping will show up in most of the bootstrap trees, because each bootstrap sample will almost always include some of the nucleotide positions that support the grouping. But if the grouping is supported only by a few nucleotide positions, then the bootstrap samples may frequently fail to include those positions, and the grouping won't show up in most of the bootstrap trees.

For example, consider the sequences in Figure 11.7: we have four sequences with three phylogenetically informative positions (positions 1, 3, and 4), and as we saw previously, Tree #1 is the best tree in that it requires four mutations at these three positions, whereas Tree #2 requires five mutations and Tree #3 requires six mutations. If we now construct a bootstrap sample for these three phylogenetically informative positions, we might find that we sample position 1 twice, position 4 once, and position 3 not at all. This data set would support Tree #1. In fact, it should be apparent that any bootstrap sample that includes a majority of positions 1 and/or 4 will support Tree #1, while only bootstrap samples with a majority of position 3 will support Tree #2.

The conventional way to represent the results of a bootstrap analysis is to add numbers to the tree built from the observed data that indicate the percentage of bootstrap trees that include each group in the observed tree (these are the bootstrap values). Bootstrap analysis is one example of resampling/permutation analyses, which (as discussed in the previous chapter) are powerful and extremely useful approaches to addressing questions about aspects of the data for which no good statistical method exists. However, a word of caution is in order in terms of interpreting the bootstrap values in a tree. As stated in the previous chapter, bootstrap analysis was originally proposed for trees for different species, with the underlying expectation that DNA sequences (or other molecular genetic data) from different species would show lots and lots of nucleotide substitutions. Hence, it is reasonable to expect that valid groupings in a tree for different species should have high bootstrap values—95% or more—as they should be supported by many nucleotide positions.

However, more recently bootstrap analysis has been applied to trees based on variation within a species, in particular, trees where the OTUs are partial or complete mtDNA sequences from different humans, and to the dismay of some the bootstrap values for such intraspecific trees often turn out to be quite low. While some would interpret this as calling into question the validity of the groupings in such trees, in my own view low bootstrap values are to be expected when the overall number of polymorphic nucleotide positions in the analysis is relatively small, as then it is inevitable that at most a few positions will support any particular group in the tree. Groups defined even by only a single nucleotide position may very well be "good" groups (in that they accurately reflect the evolutionary history of the sequences), but such groups will never receive strong support in a bootstrap analysis.

Therefore, while bootstrap analysis is useful in evaluating trees involving different species and/or very dense genetic data (e.g., thousands of polymorphic positions), when the focus of our attention is on intraspecific variation—as will often be the case in this book—there are better methods than bootstrap analysis for evaluating trees. For example, maximum likelihood analysis (discussed in the previous chapter) can also be applied to DNA sequence data. You need to specify a model of nucleotide substitution (as was done in the previous section in calculating genetic distances for DNA sequences) and then find the tree that maximizes the likelihood of observing the sequences. As mentioned in Chapter 10 (but without going into the details), likelihood ratio tests can then be used to evaluate whether or not the data fit one specified tree significantly better than another specified tree. However, maximum likelihood methods are computationally intensive, and it may not be feasible to apply such analysis to large data sets.

Another alternative is **Bayesian analysis**, named for the Reverend Thomas Bayes, who was both a Presbyterian minister and a mathematician and published exactly one work in each field during his lifetime. The accomplishment for which he is most noted, namely Bayes' theorem, was published in 1763 (2 years after his death) by an admirer, from notes that Bayes had left (Bayes and Price 1763). Bayes' theorem essentially provides a way to calculate the probability of an event from an assessment of the prior probability of the event happening, along with additional information based on the data we have obtained. The basic idea is that we should use all the information at hand, both what we have good reason to believe or already know (i.e., the prior probability), plus what we have learned by gathering additional data, in estimating the probability of an event—this is then called the posterior probability. This approach seems eminently reasonable, and indeed Bayesian analysis has been called "mathematics on top of common sense" (Malakoff 1999). For those of you not already familiar with this approach, Box 11.1 provides an example of Bayesian analysis in action.

How does Bayesian analysis work in the context of constructing a tree for DNA sequences? In theory, one would assign a prior probability to each possible tree, then assess the likelihood of the observed DNA sequence data for each tree, and then combine these to come up with the posterior probability of each tree. The tree with the highest posterior probability (or set of trees with posterior probabilities above some cutoff value) would then be the desired result. Alas, this approach fails for several reasons. First, we usually

BOX 11.1 ■ Example of How Additional (Prior) Knowledge Can Influence the Estimate of the Probability of an Outcome

The classic example of Bayesian analysis involves a hypothetical test for a disease and the probability that an individual who tests positive actually has the disease. Suppose extensive testing shows that a positive test result is obtained in 99% of the people who have the disease (and therefore 1% of the people who have the disease have a negative test result and hence are false negatives). Moreover, suppose that a positive test result is also obtained in 1% of the people who do not have the disease—these are false positives. If you then go for testing and end up with a positive test result, you may conclude that there is a 99% chance that you have the disease. That is your estimate of the probability of the outcome (having the disease) based solely on the data (the test results). But now suppose that you also know that the overall frequency of the disease in the population is 0.1%—how does this information change your chances of having the disease? While we can work this out with probability formulas, it's much easier to reason as follows: suppose we have a population of 1 million people. Then we expect about 1000 people (0.1%) in this population with the disease and hence 999,000 people without the disease. Of the 1000 people with the disease, 990 (99%) will test positive, while of the 999,000 people without the disease, 9990 (1%) will test positive. Thus, we expect a total of 990 + 9990 = 10,980 people with positive test results. And so if you have a positive test result, your chances of having the disease are 990/10,980, which is about 9%—still worrying, but not nearly as worrying as the estimate of 99%, which does not use any prior information about the incidence of the disease.

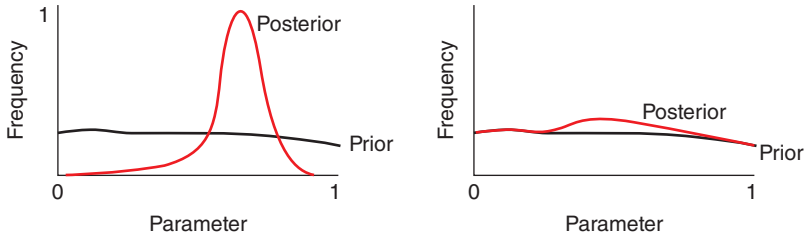


FIGURE 11.9

Examples of good (left) and not-so-good (right) posterior probability distributions for a hypothetical parameter. In this example, the prior distribution is uniform in the interval (0,1), meaning that during the analysis the hypothetical parameter can take on any value between 0 and 1 with equal probability.

have no idea how to come up with a prior probability for each tree. Second, except for very small data sets, it would take too much computational time to carry out this approach. And third, in order to calculate the likelihood of observing our DNA sequences for each tree, we also need to specify an evolutionary model for nucleotide substitutions (i.e., Jukes–Cantor, Kimura, with or without rate variation among sites, etc.)—the posterior probability of a tree thus depends not only on the prior probability of the tree and the sequence data but also on the evolutionary model used.

In practice, what one does is to assume a uniform prior probability for each tree—that is, all trees start out as being equally likely. This may seem somewhat counterintuitive in a Bayesian analysis—after all, the point is to make use of prior information—but another nice feature of Bayesian analysis is that we can compare the prior and posterior probability distributions to see how much impact the data have had on the posterior probability. Figure 11.9 shows hypothetical examples of prior and posterior probabilities for estimating a hypothetical parameter; on the left the posterior distribution differs greatly from the (uniform) prior distribution, so we could conclude that the resulting parameter estimate largely reflects the additional information coming from the data, whereas on the right the posterior distribution is not much different from the prior distribution, and hence we would conclude that our data haven't really added much to what we already knew about this parameter (which, in the case of a uniform prior distribution, is basically nothing!). It is very useful to have some indication as to how much the data have actually contributed to our estimate of the posterior probability of an event.

Having specified a prior probability distribution (usually, but not always, uniform across the range of possible values that we can reasonably expect), we next need a model as to how our sequences evolve. Fortunately, programs exist that will evaluate a set of sequences against the predictions of each potential model (based on base composition, transitions versus transversions, and rate heterogeneity) and tell you which model best fits your data, so this part is straightforward. Still, you should always keep in mind that the posterior probabilities of the trees depend not just on the priors and the data but also on the choice of model. Under one model, we might find strong support for grouping two specific OTUs together, whereas under another model, we might find strong support for other arrangements, so the choice of model can be important.

Given the prior probabilities, the evolutionary model, and the data, how does Bayesian analysis then find the trees with the highest posterior probabilities? The procedure that is currently employed is called **MCMC**, which stands for Markov Chain Monte Carlo. A precise, thorough description of MCMC and what it entails is beyond the scope of this book (indeed, it is beyond the scope of this author!), but consider the following example: suppose we have a board game, such as Monopoly, in which movement around the board is governed by the throw of two dice, and we want to know the probability of visiting each square on the board during the course of a game. While this could be worked out mathematically, it is a lot easier to simply put a playing piece on the board and start throwing the dice, keeping track of how many times you land on each square (it would be even easier to just program a computer to do this)—divide this by the total number of dice throws and that is your desired probability. After you throw the dice enough times, you will find that your probability estimates don't change much—your estimated values have converged to the true values. This is an example of an MCMC analysis: the probability that your piece ends up on a particular square after throwing the dice depends only on where your piece is now and not on how it got there, and this is what is meant by a Markov chain (the next state in a Markov chain depends only on the current state and not on any of the preceding states or how you actually got to the current state). The square that your piece does land on is governed by the throw of the dice, which introduces a random element, which is what is meant by Monte Carlo (basically, that there is

an element of randomness in how you move through the possible states in the Markov chain).

Markov chains were the development of the Russian mathematician Andrey Markov in 1906 (although they were not formally called Markov chains until 1926), and his first application of Markov chains was to calculate the probability of a vowel following either another vowel or a consonant in Pushkin's celebrated poem "Yevgeniy Onegin" (for a highly readable account of Markov and Markov chains, see Hayes 2013). Monte Carlo methods date back to 1946, when the mathematician Stanislaw Ulam was pondering how to figure out the probability of winning a game of Solitaire and decided that it would be a lot easier to simply play a lot of games and keep track of the outcomes than it would be to actually try to compute the probability (as recounted in Eckhardt 1987). Ulam was participating in the development of the atomic bomb at Los Alamos and had access to the first electronic computer, and he realized that one could make use of computers to implement Monte Carlo methods (so Solitaire has an association with computers way before Microsoft and PCs!), but somehow the first published description of Monte Carlo methods failed to mention this application and instead focused on more mundane problems in mathematical physics (Metropolis and Ulam 1949). The first union of MC (Markov Chain) with MC (Monte Carlo) was in 1953 (Metropolis et al. 1953), and while there were further developments over the years, there was limited implementation of MCMC methods until the 1990s, when computers became powerful enough to make MCMC methods feasible for problems of interest.

The way MCMC works in Bayesian phylogenetic analysis is that you start by choosing an initial tree, either at random or because you have reason to think it might be a good fit to the data (this is like the board game, where you begin on some square). You then move from the starting tree to another tree by some method that incorporates randomness, so you have some chance of arriving at every possible tree (akin to throwing the dice to move around the board). You then calculate the likelihoods associated with each tree, decide to either accept the new tree or stay with the old tree, and then repeat this process—ideally, many millions of times, keeping track of how many times each tree is chosen at each step of the chain. The more often a tree is "visited" during the chain, the higher the posterior probability associated with that tree. Sounds simple enough in principle, but the devil is in the details. In particular, you want to ensure that you adequately explore the space of all possible trees—if there are trees that are just as good or better than the ones your chain keeps visiting, but you never find them because you search only part of the space of possible trees, then you won't get a very good answer (in the board game example, this would be like using dice with only even numbers on them to move around the board, as then you'll never visit the odd-numbered squares). Another issue that frequently arises is how to ensure that your Markov chain has run long enough to give you a good estimate of the posterior probability distribution (in the board game example, if you throw the dice only a few times, you won't get a good estimate as to how frequently you land on each square). Still, MCMC methods in general are quite powerful and are seeing wide use in Bayesian phylogenetic analysis. One of the big advantages is that you get an estimate of the posterior probability of any group in the tree of interest (remember, based on not only the data but also the model of sequence evolution and the prior probability), and these posterior probabilities seem to be a better representation of the strength of support for particular groups in the tree than bootstrap values, especially for intraspecific comparisons. An example of a Bayesian MCMC tree is shown in Figure 11.10.

NETWORK ANALYSES

A drawback of using trees to represent the relationships of sequences or haplotypes (OTUs) is that there is an implied assumption that none of the ancestors are also present—in other words, all of the OTUs occur as tips in the tree. But when this is not the case—as happens frequently with intraspecific data—then trees don't do such a good job of visualizing the relationships in the data, as shown in Figure 11.11. Moreover, suppose we have more than one most parsimonious tree for the data, which also happens frequently; in such cases, a single tree does not show us all of the possible relationships in the data (e.g., see Figure 11.12).

This is where **network** analysis comes in. A network represents OTUs as nodes that are connected by links, with the length of each link proportional to the number of changes that have occurred between the OTUs that are connected by that link. Networks can be constructed from DNA sequence or RFLP data, amino acid sequence data, and short tandem repeat (STR) data—in general, if the data are binary (presence/absence of a restriction site, or DNA sequences with no more than two nucleotides at any position), then **reduced-median networks** are constructed, while if the data are multistate (e.g., STR data), then **median-joining networks** are constructed. The most common use of networks is for analysis of mtDNA sequence or Y chromosome STR data, as these are haploid, but diploid data can also be analyzed as long as the data are in the form of haplotypes (determined either experimentally, such as from family data, or computationally). The advantage of network analysis

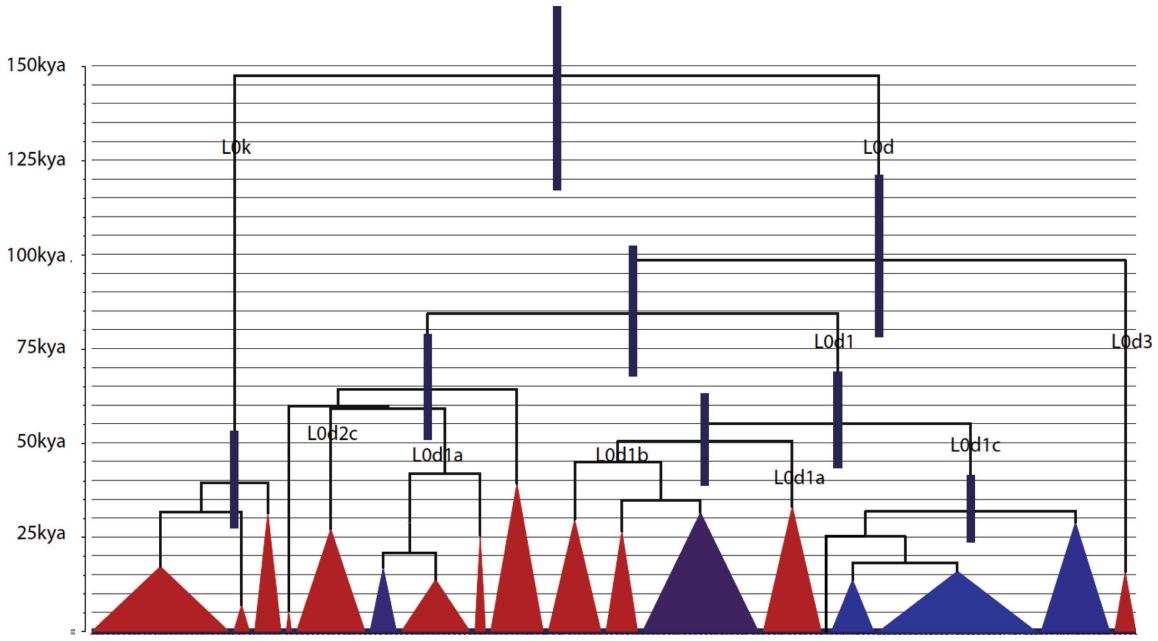


FIGURE 11.10

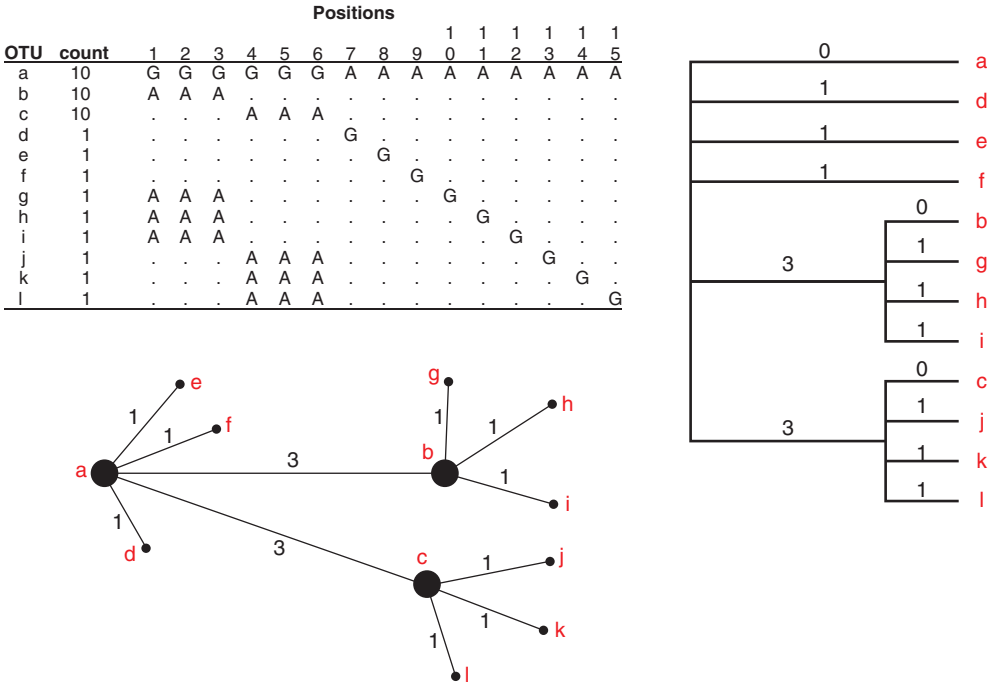
Example of Bayesian phylogenetic tree, constructed from complete mtDNA genome sequences from southern African populations. The vertical bars show the confidence intervals associated with the time estimates for each node, the width of each triangle is proportional to the number of sequences in each clade (with the corresponding haplogroup indicated), and the color of each triangle indicates the degree of support for the existence of the clade (red is high support, blue is low support). Modified with permission from Barbieri, C., et al., "Ancient substructure in early mtDNA lineages of southern Africa," *American Journal of Human Genetics* 92:285, 2013.

is that, under the appropriate conditions, the resulting network will contain all of the most parsimonious trees. Alas, as with maximum parsimony analysis, it is often not computationally feasible to construct the network with all most parsimonious trees, but there are various tricks one can then do to arrive at networks that show all of the major connections among the central OTUs (often referred to as a skeleton network), upon which the additional OTUs can be added.

Depending on the size and complexity of the data, the resulting networks may be quite "messy," with numerous links among both OTUs and hypothetical ancestors (also called **median vectors**). There are various tricks one can employ to remove alternative links, such as to prefer links between existing OTUs to those that involve median vectors (because OTUs are real data whereas median vectors are hypothetical intermediates). Another common strategy, which we already saw in the case of trees, is to give more weight to changes that are likely to be rare (i.e., transversions vs. transitions and/or slow-evolving vs. hypervariable nucleotide positions in mtDNA, or slow-evolving vs. fast-evolving STR loci on the Y chromosome). Figure 11.13 shows an example of a network before and after such processing. In sum, networks are an extremely versatile and useful way of depicting the relationships among sequences/haplotypes when both ancestors and descendants are present and/or when there are multiple most parsimonious trees for the data, and the availability of software to readily construct networks has contributed to their abundance in the scientific literature (especially, but not only, in studies of mtDNA or Y chromosome variation).

GENOME-WIDE DATA: UNSUPERVISED ANALYSES

Most of the analyses described so far in this chapter are applicable to all sorts of data (mtDNA sequences, Y chromosome STRs, etc.), as long as the individual is the unit of analysis. In this section, we focus on analyses that can be applied only to dense genome-wide data. While the amount of full genome sequence data is increasing at an accelerating pace, for the moment dense genome-wide data are most often obtained via so-called SNP chips, which typically result in genotypes at several hundred thousand to millions of SNPs across the genome per individual. However, as was discussed in Chapter 7, while SNP chips are an efficient and cost-effective way of obtaining dense genome-wide data, ascertainment bias (i.e., how SNPs were chosen for inclusion on the chips) is a serious concern, and we'll see a further example later in this chapter.

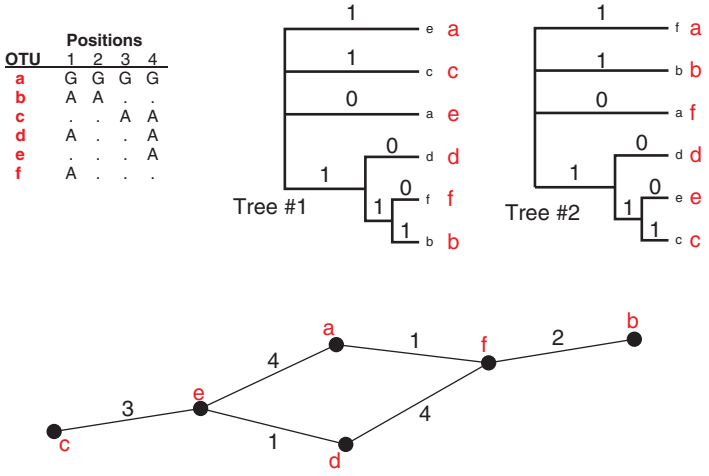


		Positions														
OTU	count	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	10	G	G	G	G	G	G	A	A	A	A	A	A	A	A	A
b	10	A	A	A	.	.	.	.	.	.	.	.	.	.	.	.
c	10	.	.	.	A	A	A	.	.	.	.	.	.	.	.	.
d	1	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
e	1	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.
f	1	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.
g	1	A	A	A	.	.	.	.	.	.	G	.	.	.	.	.
h	1	A	A	A	.	.	.	.	.	.	.	G	.	.	.	.
i	1	A	A	A	.	.	.	.	.	.	.	.	G	.	.	.
j	1	.	.	.	A	A	A	.	.	.	.	.	.	G	.	.
k	1	.	.	.	A	A	A	.	.	.	.	.	.	.	G	.
l	1	.	.	.	A	A	A	.	.	.	.	.	.	.	.	G

FIGURE 11.11

Example of DNA sequences and associated maximum parsimony tree and network, where ancestors are present in the data. "Count" indicates the number of individuals with that sequence. Numbers on each branch/link in the tree/network indicate the number of mutations that occurred along that branch/link. Note that in the standard "cladogram" version of the tree shown, branch lengths are not proportional to the amount of change, so you have to look at the number of changes on each branch to see how long the branch is and to figure out that sequence a is ancestral to sequences d, e, and f, sequence b is ancestral to sequences g, h, and i, and sequence c is ancestral to sequences j, k, and l; this information is readily apparent in the network. Note also that there is no easy way to visualize the number of individuals with each sequence in the tree, whereas in the network this information is depicted by having the size of the nodes (circles) proportional to the number of individuals with that sequence. OTU indicates operational taxonomic unit.

Many of the analyses already described in this and the previous chapter can also be applied to genome-wide data. For example, we can construct a distance matrix between each pair of individuals in our sample based on the proportion of shared alleles and then construct a neighbor-joining tree that would have each individual's multilocus genotype at the tips of the tree. But there are other analyses that can be



	Positions			
OTU	1	2	3	4
a	G	G	G	G
b	A	A	.	.
c	.	.	A	A
d	A	.	.	A
e	.	.	.	A
f	A	.	.	.

FIGURE 11.12

Sequences for which there are two equally parsimonious trees and the resulting trees and a network. The numbers on the branches in the trees indicate the number of changes that occurred on that branch, while the numbers on the links in the network indicate the position that has changed on that link. Note that it is not so easy to see which relationships are consistent across trees, whereas in the network it is quite clear which OTUs and which sites are involved in the equally parsimonious representations of the data. OTU, operational taxonomic unit.

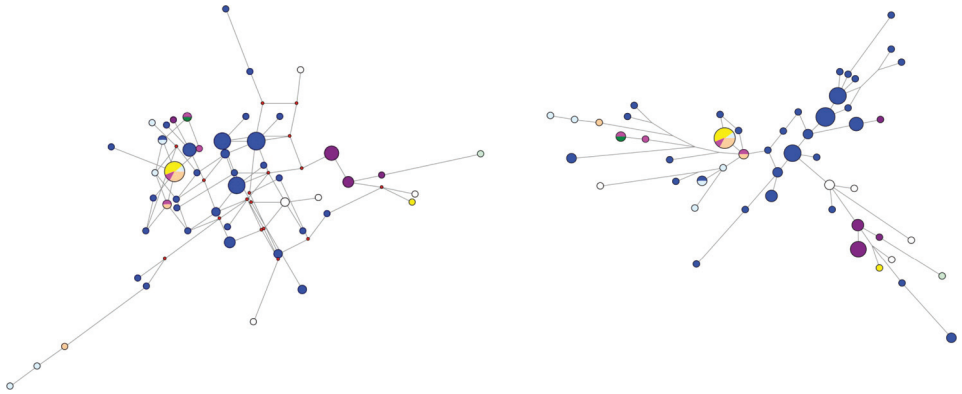


FIGURE 11.13

Example of the effect of weighting mutations in a network analysis. On the left is a network for STR haplotypes on the background of a particular Y chromosome haplogroup, with circles denoting haplotypes (the size of each circle is proportional to the frequency of the haplotype, and colors indicate different populations) and small red dots indicating hypothetical ancestors (haplotypes that are not observed in the actual data but inferred to exist). Note that there are many reticulations (alternative mutational pathways) between haplotypes in this network. On the right is a network for the same data, but weighting the data for each STR locus by the inverse of the mutation rate (so slowly evolving STR loci are given more weight, meaning that fewer mutational events are preferred at such loci). Note that there are many fewer reticulations in this network. Data used to generate these figures are from Delfin, F., et al., "The Y chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups," *European Journal of Human Genetics* 19:224, 2011.

done only on genome-wide data, and here we focus on two of these. These analyses are also known as **unsupervised analyses—**not because they are analyses done by students without the knowledge of their supervisors, but because they do not require information about population affiliation. This is an extremely important and useful aspect of the analysis of individual-level data. As we saw in the previous chapter, in a typical study samples have to be grouped into populations and then analyses are carried out on these predefined groups of samples. But if the way you group samples into populations does not correspond to the underlying genetic structure—for example, if there are subgroups within a population that differ genetically, or if there are individuals who actually should be placed into a different group—then this can have an impact on the results of the analyses.

Individual-level analyses, such as we have already seen for mtDNA sequences or Y chromosome STR haplotypes, nicely avoid this issue. You carry out your analyses at the individual level—for example, you can construct a network of the individual mtDNA sequences—and then apply group labels that you think are appropriate to see how the genetic structure indicated in the analysis corresponds to the groups (an example is shown in Figure 11.14). Unsupervised analyses extend this same principle to genome-wide data: you analyze individual multilocus genotypes, without incorporating any information about group membership in the analysis and then afterward apply the group labels.

Unsupervised Analyses: Principal Components Analysis

There are two major kinds of unsupervised analyses, and the first we have already seen applied to population-level data in the previous chapter: principal components analysis (PCA). Recall that PCA is a way of reducing the complexity of large multidimensional data sets while retaining the maximum amount of information. If we start with the underlying data, we have a matrix of up to hundreds (or even thousands) of individuals, each with up to several hundred thousand SNP genotypes. No mere mortal can gaze at such a matrix and make any sense out of it, so what PCA does is extract components that are each independent of one another, and such that the first component accounts for the largest amount of variability in the data, and each successive component accounts for less of the variability. Typically, the components (known as PCs) are plotted one against the other, with plots of PC1 versus PC2 showing the most important patterns in the data—although, as we shall see below, there are other ways of depicting the PCs.

Figure 11.15 shows just such a plot of PC1 versus PC2 for a subset of the individuals (4–5 from each population) from the Human Genome Diversity Panel (HGDP), genotyped for nearly 1 million SNPs. The HGDP was mentioned in Chapter 9, and go back and take a look at Figure 9.10 if you need to remind yourself about the populations included in the HGDP. By the way, the use of approximately equal

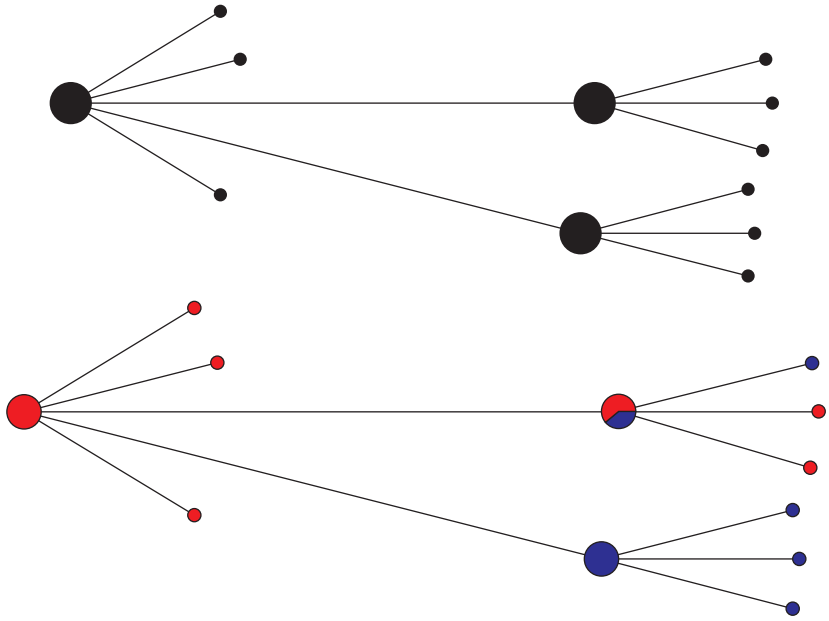


FIGURE 11.14

Top, network without group labels; bottom, the same network with labels added for two populations (red and blue), showing that one sequence is shared between the two populations while all others are unique to one population or the other.

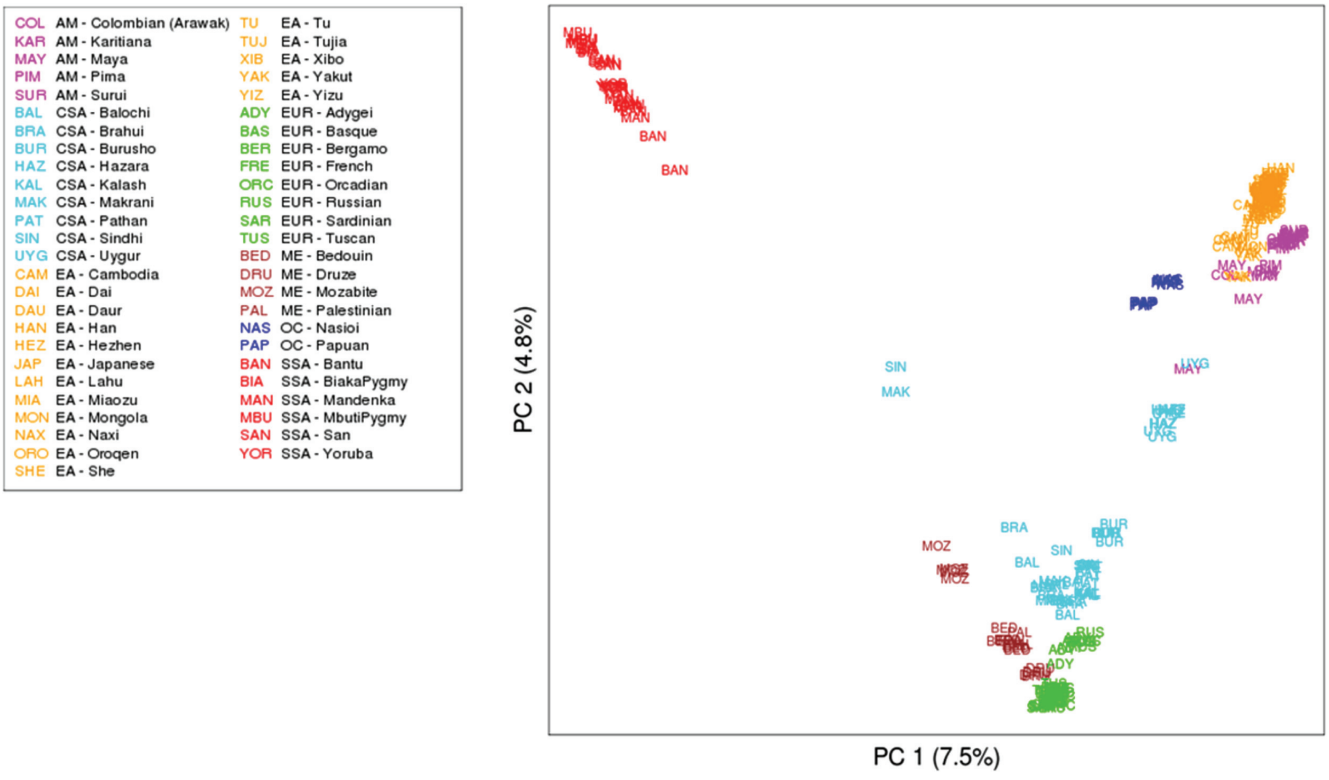


FIGURE 11.15

Plot of PC1 versus PC2 for a subset of the CEPH–HGDP samples, consisting of five individuals chosen at random from each population. Each three-letter symbol indicates the genotype (based on ~1 million SNPs) for an individual from that population, colored by continental origin according to the key to the left of the plot. Reprinted with permission from López Herráez, D., et al., "Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs," *PLoS One* 4:e7888, 2009.

sample sizes is recommended for PCA, as otherwise those groups with much larger sample sizes can distort the analysis. The individuals are labeled according to population affiliation, with different colors indicating major continental groups—although keep in mind that the population affiliation was not used to generate the plot, the labels/colors were added afterward. This plot shows a very familiar pattern, with three major vertices comprising Africa, Europe/Mideast/Central–South (CS) Asia, and East Asia/Americas/Oceania. Note that while most individuals from the same group cluster fairly closely together in the plot, there are two distinct outlier individuals, one Sindhi and one Makrani, both from Pakistan but both shifted much closer to the African groups. A reasonable hypothesis is that these individuals differ from others in their groups by having recent African ancestry. Identifying such outliers is an important result from PCA, as we would not want to include such atypical individuals in, for example, the kinds of demographic analyses discussed in the next chapter. However, it may also be of interest to explore further why outliers show up as outliers, so they shouldn't be eliminated from all further consideration—in the aforementioned example of the Sindhi and Makrani, it might be quite interesting to follow up on why these individuals have recent African ancestry.

What about other PCs—what do they show? Actually, we should first ask how many PCs are relevant, and it turns out that while the first 15 PCs are each significantly greater than zero, there is a "leveling-off" of how much of the variation is explained by around PC6 or PC7 (see Figure 11.16, left). Plotting each of these PCs is cumbersome and such plots can be difficult to interpret; an easier way to visualize the PCs is by a heat plot, in which the PC value for each individual is expressed as a color (see Figure 11.16, right), with (in this case) blue shades showing low values for the PC and red shades showing high values. Inspection of the heat plot shows: for PC1 the Africans are at one end of the scale, East Asia/Americas/Oceania at the other end, and Europe/Middle East/CS Asia in between; for PC2 the Africans are at one end, Europe/Middle East/CS Asia at the other, and East Asia/Americas/Oceania in between; PC3 distinguishes the Americas from the other groups; PC4 distinguishes Oceania from the others; for PC5 the Mbuti Pygmies and San are at one end of the scale and all other groups are at the other end, except for Biaka Pygmies who are in between (this one is intriguing as the two Pygmy groups and the

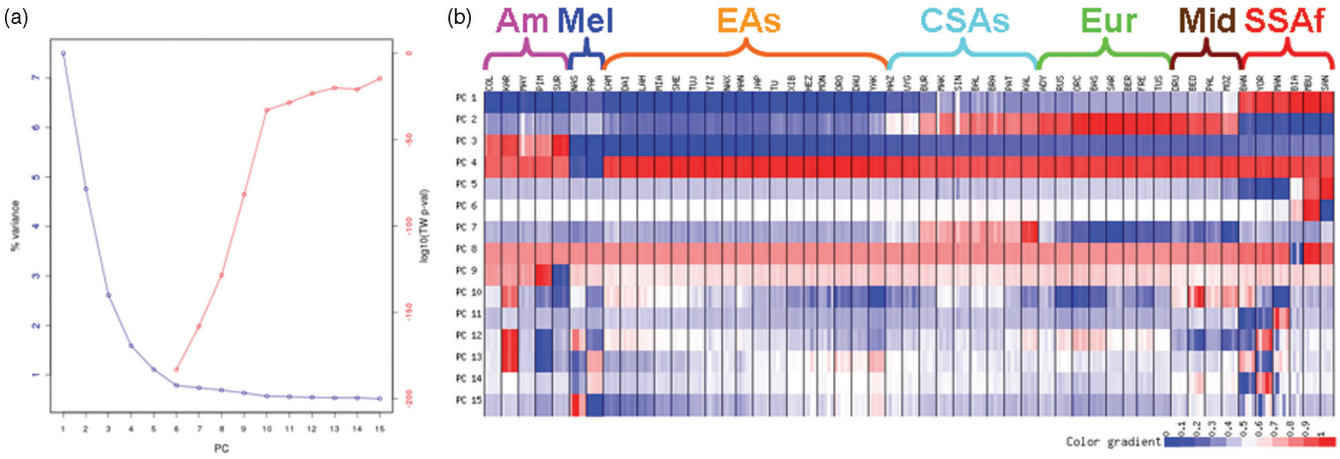


FIGURE 11.16

PC analysis of the CEPH–HGDP data. (a) Plot of the percent variation explained for each of the first 15 PCs (blue line), and a statistical test of the hypothesis that the percent variation explained by a PC is not significantly different from zero. The results indicate that the percent variation explained by each of the first 15 PCs is indeed significantly greater than zero (you'll have to trust me on this, the test is too complicated to explain here). (b) Heat plot of the value of each of the first 15 PCs for each individual genotyped in the CEPH–HGDP data. PC values have been normalized to range from 0 to 1. Each small rectangle contains five lines; each line corresponds to the value for that PC (row) for an individual from that population (column). The color of the line indicates the PC value according to the scale at bottom right. This way of visualizing PCs makes it immediately clear which populations/individuals are distinguished by each PC; for example, PC4 distinguishes the two Oceanian populations from all other populations. Continental groupings are as follows: AM, Americas; OC, Oceania; EA, East Asia; CSA, Central–South Asia; EUR, Europe; ME, Middle East; SSA, sub-Saharan Africa. Reprinted with permission from López Herráez, D., et al., "Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs," *PLoS ONE* 4:e7888, 2009.

San are traditional hunter-gatherers); and PC6 distinguishes most of the CS Asians (with the exception of the Hazars and Uyghurs) from all of the other groups. The remaining PCs distinguish either individual groups or particular individuals from everyone else.

It is tempting to interpret the above patterns in the PCs as indicative of various long-distance migrations and expansions: PC1 and PC2 would then correspond to the out-of-Africa migration, PC3 the colonization of the Americas, PC4 the colonization of Oceania, and so forth. But while this could indeed be the case, it need not be the case. The patterns in the PCs could just as easily be generated by other processes, such as genetic exchange among long-established populations mediated by geographic distances. Other analyses, described in the next chapter, are needed to distinguish among various competing hypotheses that might have produced the patterns in a PCA. The important take-home message that I will stress more than once precisely because it is both so important and yet so often ignored or abused is that PCA is extremely useful as a descriptive analysis for visualizing the dominant signals in dense genome-wide data. However, it is only a descriptive analysis and in and of itself does not tell you what sort of history generated the signals. Determining which among various competing hypotheses is most likely to lie behind the patterns in a PC plot requires the sorts of analyses described in the next chapter; claiming that the results of a PC analysis are in accordance with a particular historical explanation, in the absence of such analyses, amounts to little more than storytelling.

The global patterns of human genetic variation depicted in the PCA in Figures 11.15 and 11.16 are, as we shall see later, nothing really new. Where PCA of dense genome-wide data has had the largest impact is in distinguishing novel fine-scale genetic structure among closely related populations. For example, Figure 11.17 shows a plot of PC1 versus PC2 for 1387 Europeans genotyped at about 200,000 SNPs. Astonishingly, this plot mirrors a map of Europe almost exactly—one can discern the Iberian and Balkan peninsulas, the boot of Italy, Turkey and Cyprus, Scandinavia, the British Isles, and so forth, in the PC plot. Geography has an extraordinarily large influence on genetic structure in Europe; 90% of the individuals in the PC plot are placed within 700 km of their reported origin. To be sure, this analysis was restricted to individuals who reported that all four grandparents came from the same place; if you are like me and have ancestry from all over Europe, then where you end up in the PC plot won't have any correspondence with geography. But still, if you consider the history of Europe and how much population movement there has been during the past few centuries, then I for one would never have believed that simply having all of your ancestors from two generations ago from the same place would produce such a close fit between genes and geography.

Moreover, there is probably even more fine-scaled structure within Europe than can be seen in Figure 11.17. For example, Figure 11.18 shows a similar PC plot for Switzerland and neighboring regions. There is no one Swiss language; instead, in some parts of Switzerland, German is spoken, in some parts Italian, and in some parts French. And while there is considerable overlap, nonetheless in Figure 11.18 you can clearly see that German-speaking Swiss are genetically more similar to Germans, Italian-speaking Swiss are more similar to Italians, and French-speaking Swiss are more similar to French.

A question of considerable interest is then to what extent analyses of dense genome-wide SNP data can reveal fine-scale structure in non-European populations. After all, the SNP chips should be especially informative for European populations, as the vast majority of the SNPs included on the chips were discovered to be polymorphic in Europeans. And, as we have already seen, this ascertainment bias can have a big effect on estimates of heterozygosity, and so forth. So on the one hand, perhaps the SNP chips won't be so good at revealing fine-scale structure in non-European populations because those SNPs that would be especially informative aren't on the chips as they are not polymorphic in Europeans. On the other hand, maybe simply having data for hundreds of thousands of SNPs is sufficient to reveal fine-scale structure—even if there is ascertainment bias in how the SNPs were chosen, the allele frequency differences at such a large number of SNPs are still sufficiently informative to reveal fine-scale structure.

As someone who works mostly on non-European populations, I was quite curious to know the answer, and so together with my former postdoc and current colleague Manfred Kayser, we genotyped several populations from Oceania and southeast Asia for about 800,000 SNPs to see what we could learn. Included in the study were individuals from seven different Polynesian islands and two populations from the Southern Highlands of Papua New Guinea. The PCA of the Polynesians (Figure 11.19, left) shows that the Cook Islanders can be distinguished completely, and that there is some (albeit not complete) genetic differentiation among individuals from the other Polynesian islands. The two populations from the Southern Highlands speak different, albeit closely related, languages and come from villages that are only about 50–100 km apart (although in the highlands of New Guinea it may take several days to travel such distances!). And yet, the PCA (Figure 11.19, right) clearly distinguishes the individuals from these two populations.

Keep in mind that this is not really a fair test for the presence of fine-scale structure as we don't know

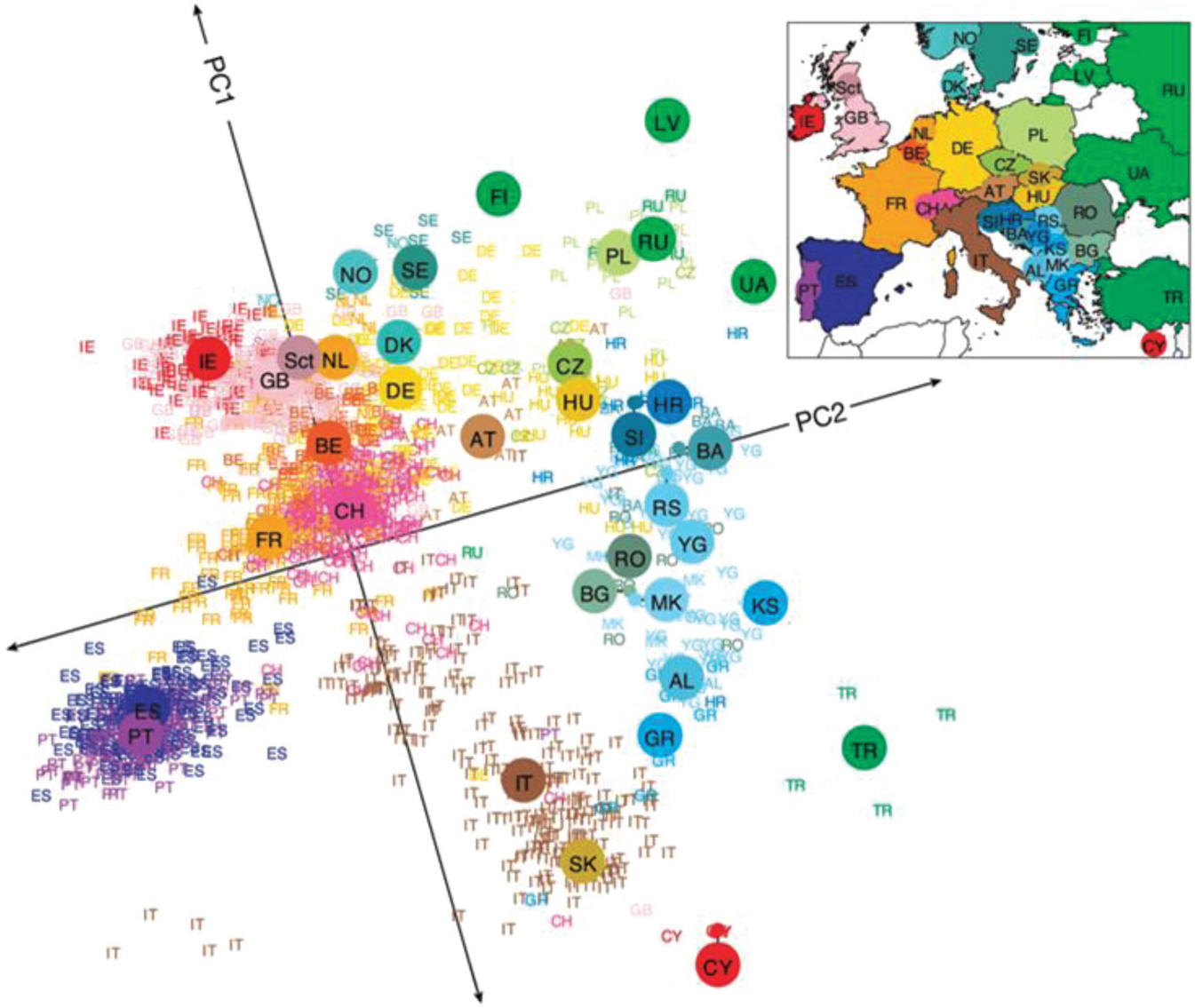


FIGURE 11.17

Plot of PC1 versus PC2 for 1387 individuals from Europe genotyped for nearly 200,000 SNPs and whose four grandparents were all born in the same town or village. The country of origin of each individual is indicated by a two-letter abbreviation, with the circles indicating the median PC values for all individuals from that country. The plot illustrates the fine-scale structure in genetic data from Europeans and has been rotated slightly to emphasize the remarkable correspondence between this genetic structure and geography: notably, the plot recapitulates a map of Europe. Reprinted with permission from Novembre, J., et al., "Genes mirror geography within Europe," *Nature* 456:98, 2008.

beforehand whether such structure really exists. Failure to see any evidence of fine-scale structure could be because such structure does not exist (e.g., individuals from different Polynesian islands really are genetically indistinguishable, or their genetic differences do not correspond to their geographic origin), or because it does exist but our analysis is not able to detect it. But, the fact that we do see genetic differences among individuals in the PC plots that correspond to geography is an indication that fine-scale structure does indeed exist and we are indeed able to detect it.

I, for one, found the results in Figure 11.19 very encouraging and, therefore, began to think that the ascertainment bias on the SNP chips was mitigated (at least, in PCA and related analyses) by simply having data from hundreds of thousands of SNPs. But a recent analysis of genome-wide SNP data from southern African populations (Pickrell et al. 2012) shows

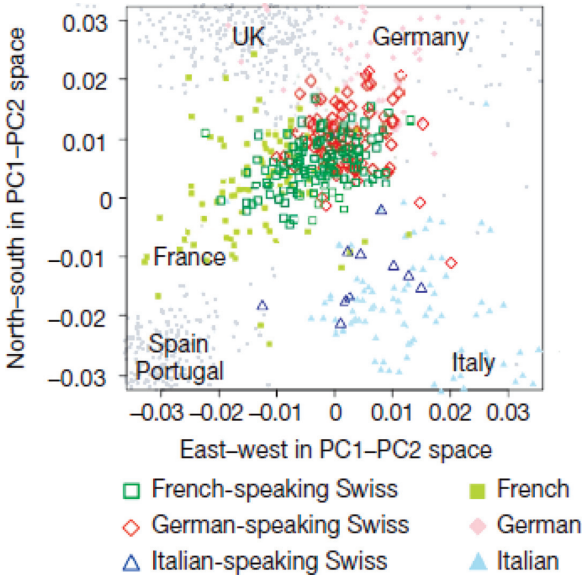


FIGURE 11.18

Plot of a subset of the data shown in Figure 11.17, focusing on individuals from Switzerland and nearby regions, showing that even at this scale there is a rough genetic correspondence between French-speaking, German-speaking, and Italian-speaking Swiss and French, Germans, and Italians, respectively. Reprinted with permission from Novembre, J., et al., "Genes mirror geography within Europe," *Nature* 456:98, 2008.

that ascertainment bias is still a reason for concern with PCA. In this study, samples were genotyped using a novel SNP chip designed by the population geneticist David Reich. This SNP chip contains several different sets of SNPs, each ascertained from one of 11 different populations from the HGDP (San, Yoruba, Mbuti Pygmy, French, Sardinian, Han Chinese, Cambodian, Mongolian, Papuan, Melanesian, Karitiana) and from two archaic humans (Neandertals and Denisovans). Each set of SNPs was chosen because they were heterozygous in a genome sequence from a single individual from that population, which thus means that the ascertainment is both simple and completely unambiguous. Figure 11.20 shows three PCA plots, all for the same southern African individuals, but for SNPs ascertained from a San, a Yoruba, or a French genome sequence, respectively. In all three plots, the first PC is quite similar and differentiates Bantu-speaking groups from "Khoisan" groups (here, "Khoisan" is used to refer to southern African groups who speak non-Bantu languages that use clicks as consonants, but neither the groups nor the languages are necessarily related). This is reassuring: the strongest signal in the data does not seem to depend on the ascertainment. However, PC2 varies dramatically in the three plots: for the San-ascertained SNPs, PC2 distinguishes among the various southern African Khoisan groups whereas the non-Khoisan groups are all the same; for the Yoruba-ascertained SNPs, PC2 distinguishes among the non-Khoisan groups whereas the Khoisan groups are all the same (note that Yoruba do not speak a Bantu language, but genetically Yoruba are closely related to western African Bantu-speaking groups); and for the French-ascertained SNPs, the Nama stand out in PC2, most likely because they have recent European admixture (although this is a hypothesis that needs further analyses to substantiate!). This is actually a cool result, because it shows that for populations that have complex histories and mixed ancestries, we can

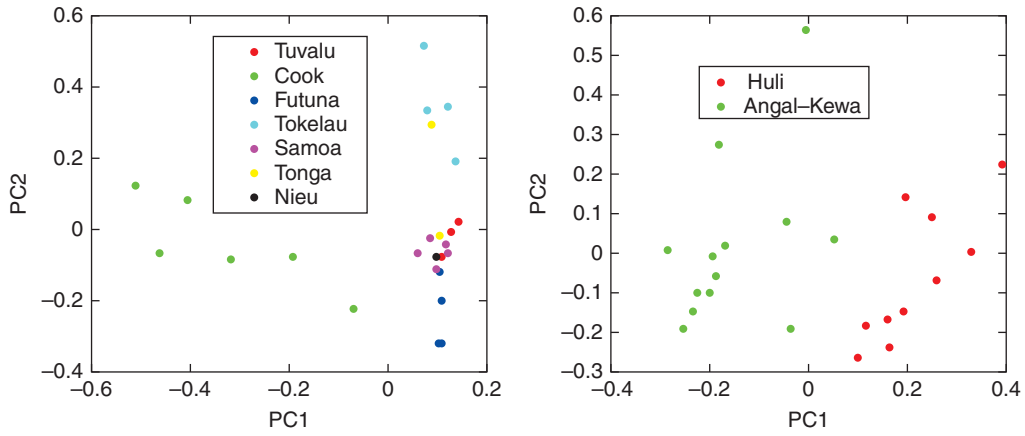


FIGURE 11.19

Plots of PC1 versus PC2 for (left) individuals from seven Polynesian islands and (right) for individuals from two different groups from the Southern Highlands of Papua New Guinea. Even with the ascertainment bias in SNP chip genotype data, it is possible to distinguish fine-scale structure in Polynesia and in highland Papua New Guinea. Reprinted with permission from Wollstein, A., et al., "Demographic history of Oceania inferred from genome-wide data," *Current Biology* 20:1983, 2010.

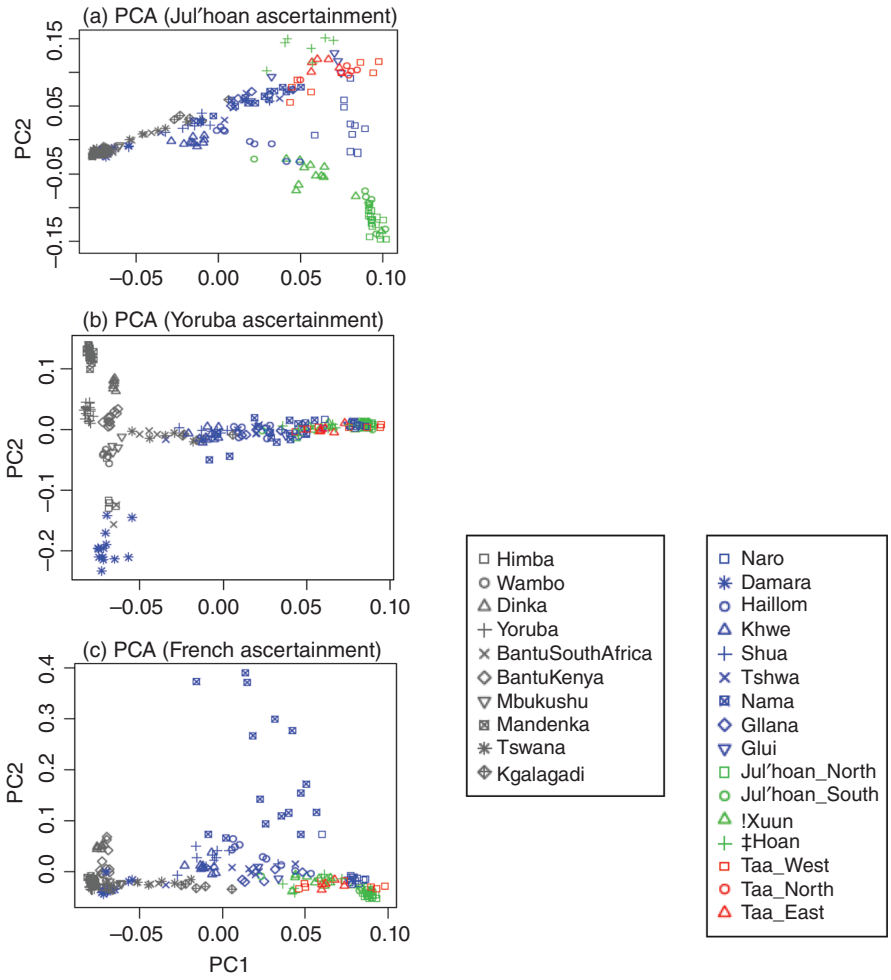


FIGURE 11.20

Plots of PC1 versus PC2 for southern Africans based on genome-wide SNPs ascertained to be heterozygous in a single (a) San (Jul'hoan), (b) Yoruba, or (c) French individual. Gray-colored symbols indicate populations speaking non-Khoisan languages, while red, green, and blue indicate populations speaking a language belonging to one of the three main Khoisan language groups. PCA, principal components analysis. Reprinted with permission from Pickrell, J., et al., "The genetic prehistory of southern Africa," *Nature Communications* 3:1143, 2012.

potentially disentangle some of this history if we can identify the genomic segments with different ancestries and analyze them separately. But the important take-home message is that while dense genome-wide data obtained from SNP chips is providing us with all sorts of interesting new insights into human population structure (and history, as discussed later), ascertainment bias does not simply disappear if one has enough SNPs—ascertainment bias always has to be considered in the interpretation of the results.

Unsupervised Analyses: STRUCTURE Analyses

The second kind of unsupervised analysis for genome-wide data is commonly referred to as a STRUCTURE analysis, after the first computer program developed by the population geneticist Jonathan Pritchard and colleagues (Pritchard et al. 2000) to carry out this analysis (although nowadays other programs tend to be used). The basic idea is that we assume that our sample of individuals comes from some fixed but unknown number of subpopulations. Each subpopulation has a different set of allele frequencies at the loci for which we have genotype data, but these allele frequencies are also unknown. What STRUCTURE analyses then attempt to do is estimate the allele frequencies in the unknown subpopulations and assign the ancestry of each individual to one or more of the subpopulations (i.e., the method allows for individuals having mixed ancestry) based on the observed multilocus genotype data. The details as to how this is done differ among the various computer programs available—STRUCTURE uses a Bayesian MCMC approach, other programs use sophisticated maximum-likelihood

approaches—and in any case are too complex for this book. Since the number of subpopulations is unknown, what one does in practice is to start by assuming that there are just two subpopulations, run the analysis, and end up with an estimate of how much of each individual's ancestry comes from each subpopulation. Then assume three subpopulations and repeat the analysis and continue in this vein, increasing the number of assumed subpopulations in each round of analysis, until the optimal number of subpopulations has been identified (in terms of how well the results fit the data—again, the details as to how this is done differ according to the method). The key point is that there is no information about any prior assumed population grouping that goes into the analysis—all that goes into the analysis are the individual multilocus genotypes, and what comes out is an estimate of the amount of ancestry that each hypothetical subpopulation contributes to each individual. So, like PCA, STRUCTURE analyses are unsupervised, in that individuals are analyzed without any use of any population labels.

To visualize the results, for each assumed number of subpopulations a bar plot is made that shows the estimated amount of ancestry that each hypothetical subpopulation contributes to each individual. These are sometimes called DISTRUCT plots, after the software that is commonly used to produce them. Even though no population information goes into the analysis, in the DISTRUCT plot the usual practice is to group together individuals from the same population/language group/geographic area/and so forth, in order to see how individuals with the same group affiliation compare in terms of their ancestry and to help identify outliers—this is analogous to using different labels to identify group affiliation for the individuals in a PCA plot, as was done, for example, in Figure 11.15.

Figure 11.21 shows the results for this type of analysis for the genome-wide SNP data from the HGDP that were analyzed via PCA in Figure 11.15. The different values of K refer to the different number of assumed subpopulations, also known as **ancestry**

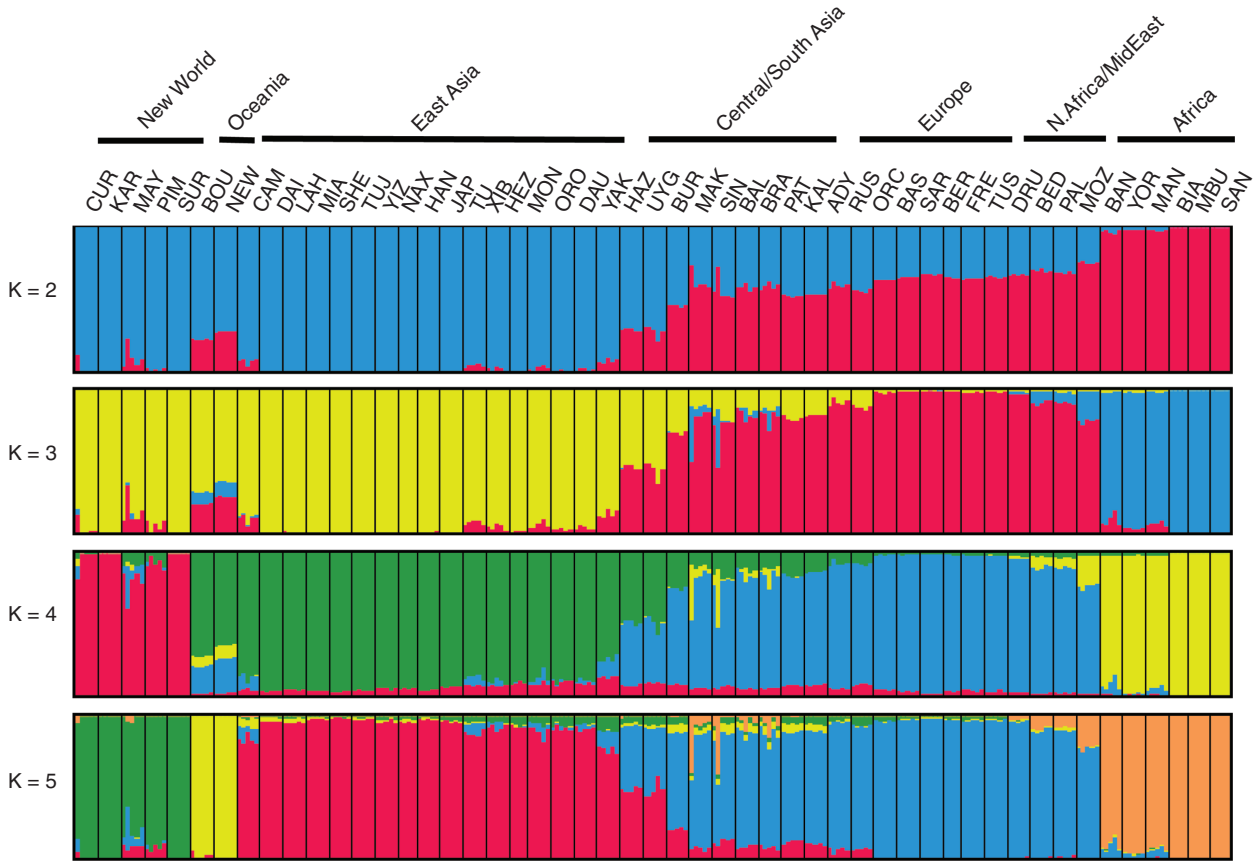


FIGURE 11.21

Results of the STRUCTURE-like analysis for the CEPH–HGDP genome-wide SNP data, for $K = 2$ to $K = 5$, with each ancestry component for a particular value of K indicated by a different color. Each rectangle contains five vertical lines, corresponding to the five individual genotypes from each population, with each line colored to represent the amount of each ancestry component inferred for that individual for that value of K. See text for further explanation. Reprinted with permission from López Herráez, D., et al., "Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs," *PLoS ONE* 4:e7888, 2009.

components. For $K = 2$, the ancestry components largely distinguish African from non-African populations. For $K = 3$, the non-African populations are further divided into European (and nearby, such as Middle East) populations and East Asian populations (and related groups, such as native Americans and Oceanians). At $K = 4$, the native Americans become distinguished, while at $K = 5$, the Oceanians are distinguished. There is thus a rough correspondence in the patterns revealed by the higher PCs (Figure 11.16) and the increasing number of ancestry components, which is a good indication that both analyses are detecting the same major signals of genetic structure and relationships in the data.

The optimal results occurred at $K = 6$, and in Figure 11.22 the results for $K = 6$ are plotted on a map, in order to help visualize how the ancestry components correspond to geography. It is readily apparent that there is a strong geographic influence on the ancestry components, with different components predominating in the Americas, Africa, Europe/Middle East/North Africa, CS Asia, East Asia, and Oceania. It is also readily apparent that some populations have mixed ancestry, for example, the Hazara and the Uyghurs. Moreover, as was seen in the PCA, one Makrani and one Sindhi appear to differ from other members of these groups in having high frequencies of the African ancestry component, suggesting (but not proving!) recent African admixture in these two individuals. Thus, as with PCA, STRUCTURE analyses are an extremely useful way of depicting underlying genetic structure, suggesting possible admixture, and identifying potential outliers. In fact, both PCA and STRUCTURE analyses are standard tools in the repertoire of methods available to analyze genome-wide data to the point that it would be highly unusual to not carry out both analyses in a study involving such data.

However, some caveats are in order. First, the results of a STRUCTURE analysis are only as meaningful as the underlying model, which assumes a fixed number of discrete subpopulations that contribute to the ancestry of each individual. If this model is wrong—and there is no good way to test it—then the results may be completely meaningless. Second, one always

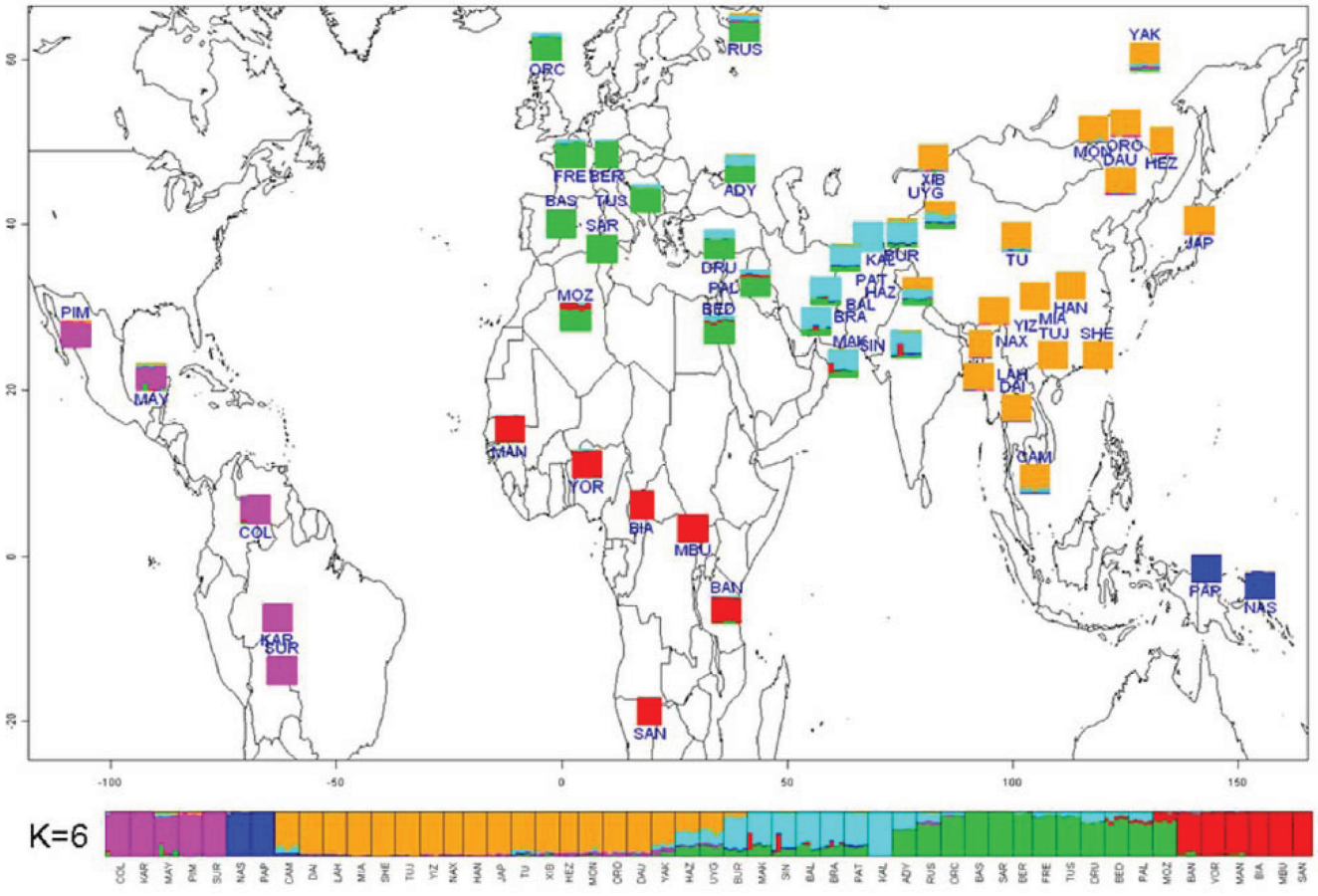


FIGURE 11.22

Results of the STRUCTURE-like analysis for $K = 6$ for the CEPH–HGDP data, with the results (below) plotted on a map (above). Reprinted with permission from López Herráez, D., et al., "Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs," *PLoS One* 4:e7888, 2009.

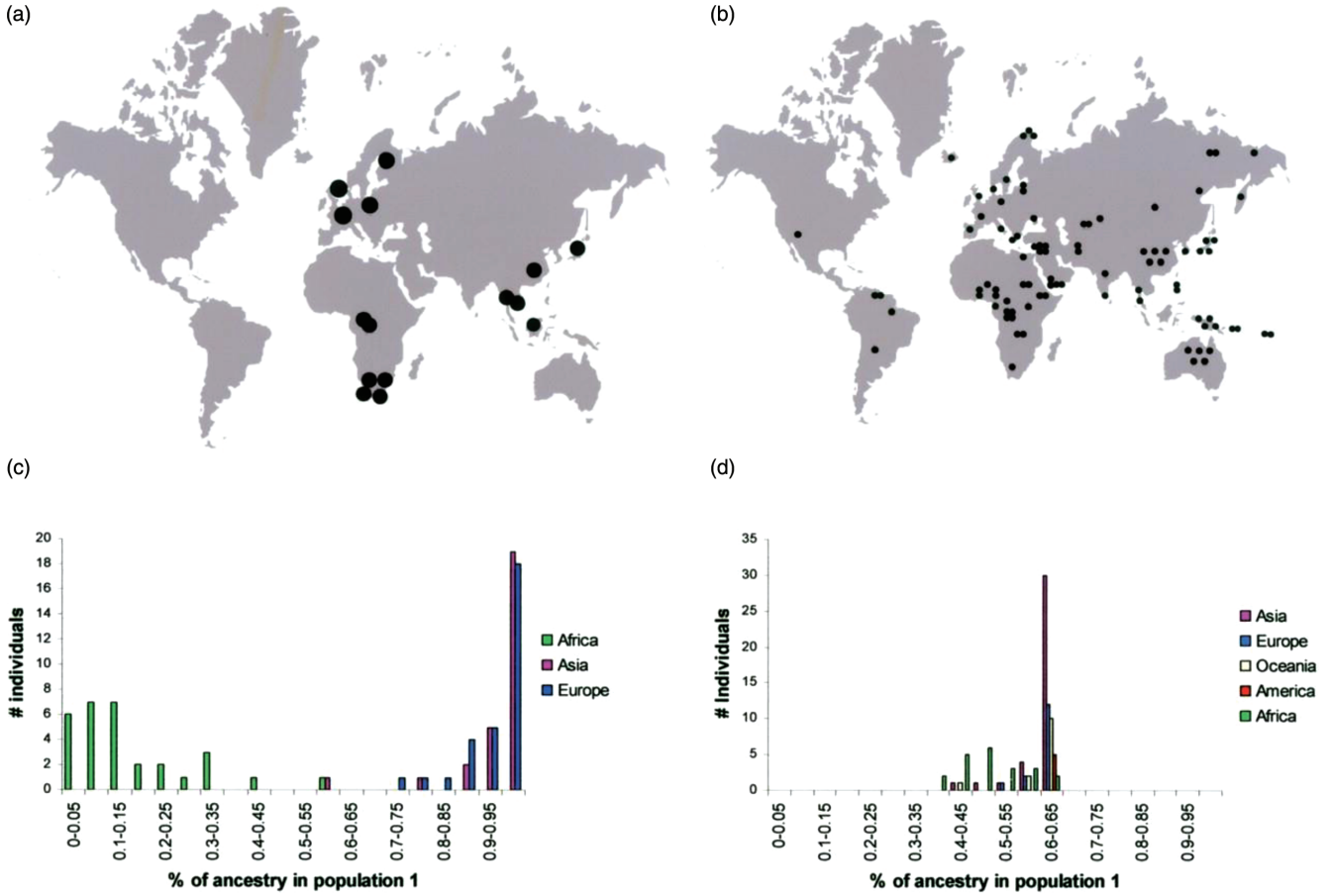


FIGURE 11.23

Illustration of the effects of sampling on a STRUCTURE-like analysis. (a) A map of sampling locations for sampling based on well-defined populations (with the size of each circle proportional to the number of individuals). (b) A map of sampling locations based on a more even geographic sampling (albeit fewer individuals per location than in (a)). (c) The assignment of ancestry to two inferred ancestral populations for the individuals in (a), based on 20 autosomal STR loci. Individuals of African ancestry are clearly distinguished from individuals of non-African ancestry. (d) The same analysis for assignment of ancestry to two inferred ancestral populations for the individuals in (b), genotyped at the same 20 STR loci. Now there is no clear distinction of African versus non-African individuals. Reprinted with permission from Serre D., and Pääbo, S., "Evidence for gradients of human genetic diversity within and among continents," *Genome Research* 14:1679, 2004.

has to consider how individuals were sampled and how this might influence the analysis. Some of you may look at Figure 11.22 and conclude that human genetic variation is organized into discrete clusters—that is, "races." However, if you confine the sampling to well-separated, discrete populations, then you should not be too surprised to find that you can readily classify individuals into populations based on their ancestry (e.g., Figure 8.1). A different sampling scheme may give you completely different results, as shown in Figure 11.23. With the sampling scheme on the left in the figure (consisting of a few locations each in Africa, Europe, and East Asia) and assuming two ancestral subpopulations, the Africans are fairly readily distinguished from the non-Africans. However, with a different sampling scheme (Figure 11.23, right) that sampled more widely across each geographic region, there is no clear-cut distinction between Africans and non-Africans. It is always important to pay attention to the sampling scheme and how it might influence the results that are obtained. As to whether or not genetic analyses support the existence of human races is a question we will discuss in Chapter 14.

Finally, as with PCA, STRUCTURE analyses help us to visualize only the underlying genetic structure of our samples; they do not inform us about the historical processes that gave rise to this structure, no matter how tempting it is to speculate about such processes. Although this point was made in the discussion about PCA, it is worth repeating here. Take, for example,

the Uyghurs in Figure 11.22, who have appreciable frequencies of the East Asian, European/Middle East/North Africa (call this European for short), and CS Asian ancestry components. There are a variety of scenarios that could produce these results:

1. The Uyghurs are descended from a population with one ancestry component (e.g., East Asia) that later admixed with two separate populations, each with one of the other ancestry components (Europe, followed by CS Asia or vice versa in this case). Note that there are several possible scenarios of this kind that involve two separate admixture events.
2. The Uyghurs descend from a population with one ancestry component as above (e.g., East Asia) and subsequently admixed with a single population that carried both of the other ancestry components (Europe and CS Asia in this case). Note that there are several possible scenarios that invoke one admixture event of this kind.
3. The Uyghurs are descended from a population with mixed ancestry involving two ancestry components (e.g., East Asia and CS Asia) and later admixed with a single population that contributed the third ancestry component (Europe). Again, there are several possible scenarios of this kind.
4. The Uyghurs are descended from a population that had all three ancestry components and thus have always had all three during their history and did not experience any subsequent admixture.

Thus, there are many different historical processes that would give rise to the same results in the STRUCTURE (or PCA) results. In particular, the presence of the same ancestry component in two different populations could happen because they are descended from a population with that ancestry component, or because migration from one population to the other (or in both directions, or even from some other population) brought that ancestry component to both populations. Unfortunately, telling stories based on STRUCTURE analyses is all too common in the literature; it isn't hard to find papers that claim that a particular ancestry component reveals a particular migration that happened at a particular time. The take-home message: STRUCTURE and PCA analyses are extremely useful in deciphering and visualizing the major patterns present in genome-wide data, but in and of themselves they do not inform you about the specific historical processes that gave rise to the observed genetic patterns. Instead, they can provide hypotheses about the historical processes that gave rise to these patterns that can be investigated further, and how this further investigation is done is the subject of the next chapter.

CHAPTER 12

INFERENCES ABOUT DEMOGRAPHIC HISTORY

In the previous two chapters, we covered the analysis of genetic data where the unit of analysis is the population and the individual, respectively. These analyses are all descriptive in that they while they provide insights into the patterns of genetic variation that characterize populations/individuals, they do not tell you what historical processes may have given rise to the observed patterns. To be sure, descriptive analyses can suggest potential scenarios—for example, a population that has low levels of genetic diversity and large genetic distances with other populations is likely to have had a small effective population size, possibly as a consequence of a bottleneck or founder event. But if we want to get beyond suggestions and speculations, we need more than descriptive analyses; this chapter will cover analyses that attempt to answer questions about the actual **demographic history** of populations and species, such as:

When did these two species diverge?
How old is the variation in this gene?
How old is this mutation?
Where did this mutation originate?
When did these two populations diverge from a common ancestor?
How has the size of this population changed over time?
How much migration has there been from population X into population Y?
When did the migration from population X into population Y occur?

It was not all that long ago that the best we could hope to obtain from genetic data would be an estimate of when two species or populations diverged (Figure 12.1). However, thanks to both more extensive genetic data (discussed in previous chapters) and novel methods for analyzing such data (covered in this chapter), much more complex (and hence realistic) demographic scenarios can be explored.

DATING EVENTS

As indicated by the aforementioned questions, we are often interested in knowing when something happened in the past. This sort of information is useful for comparing inferences from genetic data to inferences from archaeological, paleontological, linguistic, environmental, biogeographic, climatic, or other data. Moreover, knowing when something happened may help shed some light on why it happened. In this section, we'll discuss the dating of species divergence; most recent common ancestry; specific mutations; and population divergence. In a later section, we'll discuss dating admixture events.

Species Divergence Times

The first use of molecular dating approaches was to estimate species divergence time, and this still remains a very common (and powerful) application. The idea is very simple: we estimate the number of mutational differences between two species and use an estimate of the rate at which such mutational differences arise to figure out how much time it took for the observed number of mutational differences to arise. For example, if we observe a sequence divergence of 10% between two species for a particular segment of the genome, and if we know that the rate of evolution of this genomic segment is 2% per million years, then the estimated divergence time of these two species is 5 million years ago. Very simple and elegant, with no need for a fossil record or any other information about these two species.

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

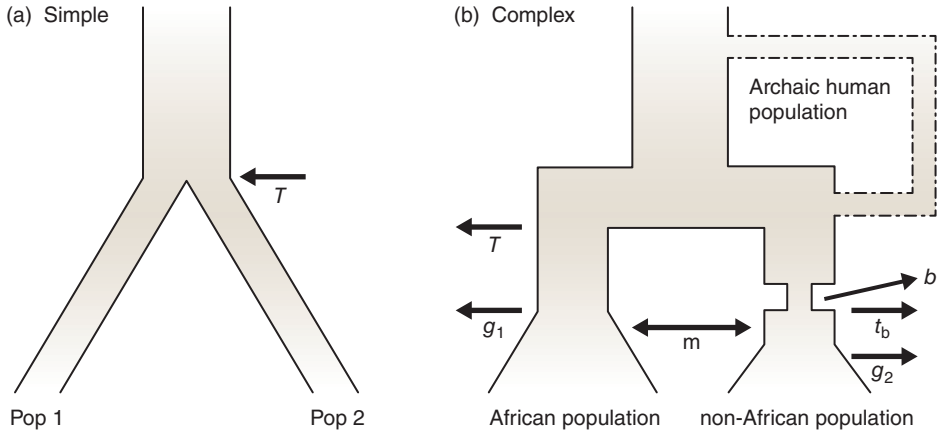


FIGURE 12.1

Illustration of simple (a) versus complex (b) models of demographic history. T is divergence time, g_1 and g_2 are periods of population growth, b is a bottleneck (with t_b the time of the bottleneck), and m is migration. Reprinted with permission from Stoneking, M., and Krause, J., "Learning about human population history from ancient and modern genomes," *Nature Reviews Genetics* 12:603, 2011.

Molecular dating thus requires two pieces of knowledge (the amount of genetic difference and the rate of genetic change over time) and an assumption (that the rate of genetic change has indeed been constant over time); we'll consider each of these. In the previous chapter, we have already discussed how to obtain appropriate estimates of the number of mutations (more properly, amount of sequence divergence) from various kinds of genetic data; the crucial point here is that one has to use the correct evolutionary model (e.g., appropriate transition/transversion ratio, whether or not all sites have the same rate of evolution, stepwise mutation model for STR loci, etc.) in order to get an accurate estimate of the number of mutations. As we saw in the previous chapter, the observed sequence divergence in the mtDNA control region of about 15% between humans and chimpanzees translates to an estimated whopping 75% sequence divergence that actually occurred, according to the evolutionary model that best fits the data. Still, getting the estimated sequence divergence between two species of interest is relatively straightforward.

Less straightforward is how one goes about estimating the rate of molecular evolution—that is, how fast the clock is ticking. The simplest approach is to have (at least) one calibration point, consisting of an estimate of nucleotide divergence (or other genetic distance appropriate to the data) for a pair of species for which there is a good estimate of the divergence time from fossil or biogeographic data. For example, a commonly used calibration point for primate studies is the divergence between the ancestors of Old World monkeys and apes. The earliest fossils confidently assigned to an Old World monkey ancestor (and not a common ancestor of Old World monkeys and apes) are at least 21 million years old, while the earliest fossils confidently assigned specifically to an ape ancestor are about 25 million years old (Begun 2007). The divergence between Old World monkeys and apes is then set at the oldest date for the unequivocal appearance of either in the fossil record, namely, 25 million years ago. You can see the potential problem: with a new fossil discovery (say, an ape or Old World monkey fossil dated to 30 million years ago) or reinterpretation/redating of existing fossils (e.g., if the fossils older than 20 million years ago are shown to be a common ancestor of both Old World monkeys and apes, or are redated to 15 million years ago), the old calibration point goes out the window. Still, a reasonably complete and accurately dated fossil record currently provides the best source of calibration points for molecular dating—and this is why it is not quite correct to call molecular dating "fossil-free dating" (as some have done), because you do need the fossil record for calibrating the clock.

Biogeographic evidence can also provide a calibration point—for example, if uplift of a mountain range divides the range of a single species, resulting in reproductive isolation and ultimately the formation of two species, then geological dating of the uplift combined with the sequence divergence of the two species can provide a calibration point. However, one has to be careful with this approach, as the divergence time between the two species may not necessarily correspond to the geological event. For example, it might seem reasonable to assume that the divergence between Old World and New World monkeys would have occurred when the African and South American continents split apart and, therefore, this event could be used as a calibration point. However, the continents

split apart around 100 million years ago, while genetic evidence based on a variety of calibration points as well as fossil evidence indicate that the divergence between Old World and New World monkeys is on the order of only 35–40 million years ago or so (Schrago et al. 2013). Since the evidence is rather overwhelming that monkeys evolved in Africa, how did they then get from Africa to South America? While rafting has been suggested as one possibility, the distance that would have been involved (on the order of 1000 km) makes rafting rather problematic. Intriguingly, there is some evidence to suggest that around 30–40 million years ago there may have been a chain of islands between South America and Africa, some of them quite substantial in size, that would have facilitated travel via short distance rafting and island hopping (de Oliveira et al. 2009).

Rates of molecular evolution based on such fossil/biogeographic calibration points are sometimes referred to as **phylogenetic rates**. Another way to estimate the rate of molecular evolution comes from family studies: by comparing DNA sequences (or STR profiles, or other genetic data) in parents and offspring, you can get a direct estimate of the mutation rate per generation. To the consternation of some, these **pedigree rates** tend to be higher than the phylogenetic rate estimates (at least, in the case of mtDNA and the Y chromosome—current pedigree estimates for genome sequences are lower than phylogenetic rate estimates, which will be discussed in Chapter 13); however, it is easy to understand why this should be the case. Recall that back in Chapter 6, we saw that the rate of neutral molecular evolution is the product of the rate at which new neutral mutations arise and the rate at which they are fixed, and overall is expected to be equal to the neutral mutation rate. But not all of the new mutations that arise in families each generation are neutral; some (unknown) fraction will be deleterious and hence eliminated by selection, while others will be lost by drift. So, we actually expect the pedigree rate to overestimate the true rate of molecular evolution.

Given all of these caveats about calibration points and the resulting estimates of the rate of molecular evolution, the best advice is to not rely on a single calibration point but rather use several different calibration points if possible. Moreover, one can use a Bayesian approach to incorporate uncertainty either in the time of a calibration point or in an estimate of the rate of evolution by using a range of values rather than a single point estimate—this can greatly improve the resulting estimate of divergence time.

The final point to consider about molecular clocks and dating species divergence times has to do with the assumption that the rate of molecular evolution has been constant over time. Critics are quick to point out that if the rate has changed over time, then the dates derived from molecular clocks may be meaningless. Fortunately, there are straightforward ways to test whether the observed data conform to the expectations of a molecular clock—in other words, the molecular clock hypothesis is a testable hypothesis and not merely an untestable assumption that one has to invoke. The first such test was devised by Vincent Sarich and Allan Wilson in the late 1960s (Sarich and Wilson 1967b), in response to vociferous criticism of their use of the molecular clock approach to date the divergence times of various ape species (including humans), which we will discuss in more detail in Chapter 13. Contemporary tests of the molecular clock assumption are more sophisticated than that of Sarich and Wilson, but the underlying principle is still much the same, so for simplicity we'll go through their version. Consider the example in Figure 12.2, where we have two species, A and B, from which we have genetic data and we want to estimate their divergence time, so we want to know whether the rate of molecular evolution has been the same along the A and B lineages since they diverged from a common ancestor. To apply the **relative rate test** of Sarich and Wilson, you need comparable genetic data from a third species, C, which you know is an outgroup with respect to A and B (i.e., you know for a fact that A and B share a more recent common ancestor with each other than either does with C). The test is then quite simple: you compare the genetic distance from A to C to the genetic distance from B to C. Note that the genetic distance from A to C consists of the amount of genetic change from the AB common ancestor to A, plus the amount of genetic change from the AB ancestor to C (see Figure 12.2).

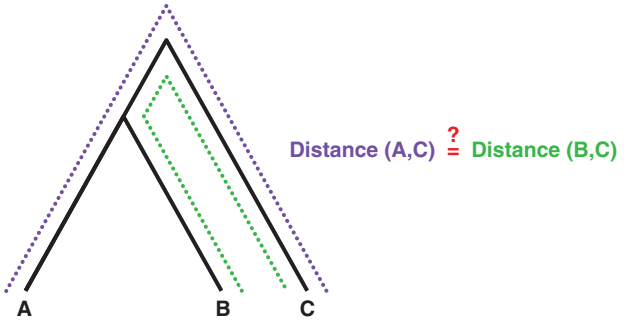


FIGURE 12.2

Relative rate test. To test whether the rate of molecular evolution has been the same for lineages A and B, compare their genetic distances to an outgroup, C. If the genetic distance between A and C (purple) equals the genetic distance between B and C (green), then the rate of molecular evolution is the same along lineages A and B. If the genetic distances are significantly different, then the rate of molecular evolution is different for lineage A versus lineage B.

Similarly, the genetic distance from B to C consists of the amount of genetic change from the AB common ancestor to B, plus the amount of genetic change from the AB ancestor to C. Thus, the A-C distance and the B-C distance both include the amount of genetic change from the AB ancestor to C; they differ only in that the former includes the genetic change from the AB ancestor to A, while the latter includes the genetic change from the AB ancestor to B. So, if the A-C distance is equal to the B-C distance, then the rates of molecular evolution along the A and B lineages since they diverged from the AB ancestor are also equal. But if the A-C distance is significantly different from the B-C distance, then the rates of molecular evolution along the two lineages are also significantly different.

It is, therefore, standard practice to carry out a relative rate test (or a more sophisticated variant thereof) before using a molecular clock approach to date divergence times. If the relevant genetic distances are equal, then we are justified in using the molecular clock (or, more formally, if the relevant genetic distances are not significantly different, then we do not reject the null hypothesis of a constant rate of genetic change). But what happens if the relevant genetic distances are significantly different and thus the data reject the hypothesis of a constant rate of change? The prudent course would be to stop there and not attempt to date the divergence time from the genetic data. The more daring course would be to employ a **relaxed molecular clock** approach, in which one allows for rate variation in different lineages while simultaneously estimating divergence times. Relaxed clocks are seeing more and more use, and there are several methods for estimating divergence times with relaxed clocks. How well they work is still a matter of discussion and investigation and depends to a large extent on how one models the rate variation—Are rates for different lineages correlated or uncorrelated? Do the rates evolve according to a particular model, and so forth? The best advice at this time would be to take any divergence time estimates derived from relaxed molecular clocks with a grain of salt, at least until we have a better idea as to how well relaxed molecular clocks actually work.

One more important point about molecular clocks: as we shall see in subsequent chapters, molecular clock dates are always accompanied by rather large variances and confidence intervals, much larger than those associated with other forms of dating such as radiocarbon dating. This is because the molecular clock does not "tick" at an absolutely regular rate, like a real clock; mutations do not occur at a fixed rate over time, as with the decay of radioactivity. Instead, the molecular clock is a stochastic clock, meaning that while mutations may occur at an average rate that is constant over time, in any given time interval there may be more or fewer mutations than expected, just by chance. It's as if our real clock had an average rate of 60 minutes per hour, but some hours have more than 60 minutes (like the hours you spend listening to a dull lecture) and some have less (like the hours you spend engaging in an enjoyable activity, such as reading this book); over the course of many many hours, there would be an average of 60 minutes per hour.

Moreover, the variance associated with molecular clock dates tends to be quite large because there are many different sources of variance in such estimates. The rate of molecular evolution is not known as a fixed constant (unlike a real clock, where we know the rate is a fixed 60 minutes per hour) but instead has to be estimated from data, and so there is a variance associated with the rate estimate. And to estimate genetic distances, we typically take a sample of individuals from a population or species, rather than sampling everyone, so there is a variance associated with the sampling of individuals—if we took a different sample, we would get a (somewhat) different estimate from the data. The larger the sample size, the smaller this source of variance, and if we sampled everyone from the population, we would eliminate this variance. Similarly, we usually sample only a portion of the genome of an individual, and not the entire genome, and that also introduces a variance into our estimate—if we sampled a different portion of the genome, we would get a (somewhat) different estimate. The more of the genome studied, the smaller this source of variance, and to eliminate this variance, we would have to sequence the entire genome from our sample of individuals (which is rapidly becoming feasible).

But even if we sequenced the entire genome from everyone on the planet, there would still be a source of variance in our genetic distance and molecular clock estimates, and that is because evolution is inherently a stochastic process. If we were to repeat the history of life on this planet, even with the same sequence of speciation events and so forth, we would have a different number of mutations in our populations and species, so this is also a source of variance in our estimates. How large this variance might be is difficult to estimate as we cannot simply rerun the evolutionary history of life on this planet many times and measure the outcome. But we can get some idea by examining the variance in estimates from independent segments of the genome (i.e., those that are not closely linked), since independent segments of the genome are assumed to have independent evolutionary histories (although, strictly speaking, this is not correct as correlations can still exist among independent segments). And this variance among independent segments of the genome tends to be quite large, much larger than the variance associated with the sampling of individuals or of the genome. So, because of the inherent stochastic nature of the evolutionary process (i.e., the

occurrence and fixation of mutations over time), there will inevitably be large variances and confidence intervals associated with molecular dates.

This section has focused on issues and potential problems that arise with the use of molecular clocks, which runs the risk of giving you the impression that these outweigh the benefits of dating species divergence times with molecular clocks. Nothing could be further from the truth: as we shall see in subsequent chapters, molecular clock dating is responsible for some of the most important insights that molecular anthropology has contributed to the field of human evolution. The fact that one can date events simply from the amount of genetic divergence, without any need for appropriate fossils or any other evidence, is indeed a powerful and useful approach. Still, in order to properly appreciate the contributions of dating via molecular clocks, one needs to understand both the strengths and limitations of the approach.

Time to the Most Recent Common Ancestor

The molecular clock approach to dating a species divergence time is one example of what is more generally known as dating the **t**ime to the **m**ost **r**ecent **c**ommon **a**ncestor (or **TMRCA** for short). The idea is that any given species must share a common ancestor with any other species, which must be true if all living things are descended from a single origin of life on this planet. The amount of genetic change between two species then provides an estimate of when they diverged from this common ancestor. However, it is important to note that the genetic divergence for a genomic segment does not necessarily correspond to the species divergence. As depicted in Figure 12.3, it is reasonable to assume that the common ancestor for two species would have been polymorphic for the genomic segment analyzed—practically all genomic segments are polymorphic in species today, so undoubtedly they would also have been polymorphic in the past. With such polymorphism in the common ancestor, then it follows that the TMRCA (or genetic divergence) estimated from a molecular clock approach is older than the actual species divergence. How much older depends on the effective population size of the common ancestor—the bigger the effective population size, the more polymorphism one expects, and hence the greater the discrepancy between the TMRCA and the species divergence time. In practice, for effective sizes commonly found in primates, such ancestral polymorphism is expected to be on the order of a few hundred thousand to at most a few million years old. These timescales are not important for species divergence times on the order of many million years but could have an impact on more recent divergence times. Thus,

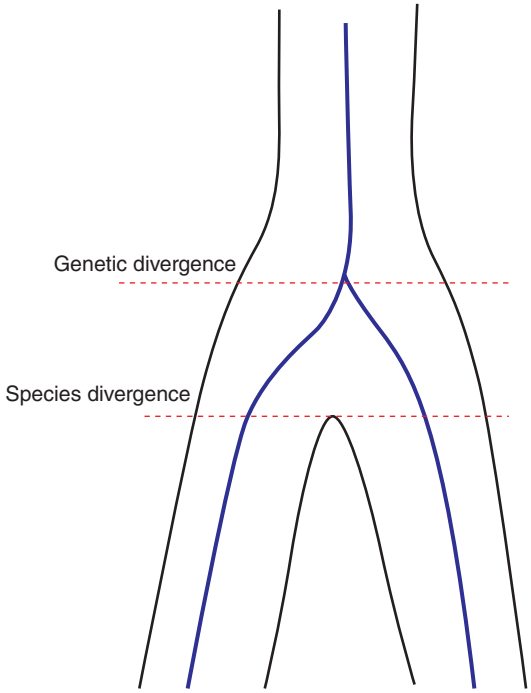


FIGURE 12.3

Diagram illustrating that genetic divergence (or coalescence) is expected to be older than the species (or population) divergence.

it is important to keep in mind that the species divergence times from molecular clock estimates provide an upper bound; the actual species divergence time is expected to be more recent.

Another important consequence of this ancestral polymorphism is that it can sometimes lead to discrepancies between the species phylogeny and the **gene genealogy** (i.e., the phylogenetic history of the genomic segments). This is depicted in Figure 12.4; ordinarily, we might expect all of the genetic variation within a species to have an MRCA (most recent common ancestor) that is also within that species (see the left panel of Figure 12.4). However, with ancestral polymorphism, it is possible to have an allele within a species A that is more closely related to an allele in another species B than it is to the other alleles in species A (see the right panel of Figure 12.4). This phenomenon is known as **incomplete lineage sorting**, and numerous examples are known. For example, even though at the species level humans are most closely related to chimpanzees, and gorillas are our next closest living relative (as we shall see in Chapter 13), for about 30% of the genome either the human sequence or the chimpanzee sequence is more closely related to the gorilla sequence than is the human sequence to the chimpanzee sequence; in other words, humans and chimpanzees are most closely related for

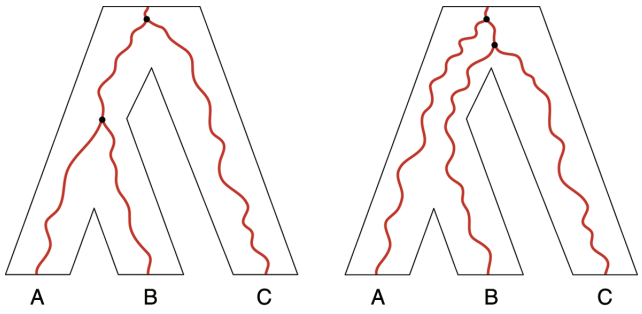


FIGURE 12.4

Incomplete lineage sorting due to ancestral polymorphism. The tree illustrates the true species relationship between species A, B, and C. On the left, the red line shows the history of a particular gene, with mutations (black circles) occurring in correspondence with the species relationships. On the right, mutations occurring in the ancestor of species A, B, and C end up, by chance, sorting into these species such that the alleles in B and C are most closely related, even though species A and B are most closely related to one another.

about 70% of the genome (Patterson et al. 2006). The extent to which incomplete lineage sorting occurs is a function of both the effective population size of the ancestral populations and how closely the speciation times are spaced; as we shall see later, the rather extensive incomplete lineage sorting for humans, chimpanzees, and gorillas is a reflection of their overall close evolutionary relationship.

We can easily extend the concept of a TMRCA to dating the origin of the variation within a genomic segment within a single species. Just as all living things descend from a single origin of life on this planet and thus are related via common ancestors, it also follows that all of the alleles of a genomic segment must trace back to a single common ancestral copy of that genomic segment. If this is not immediately obvious to you, consider the example in Figure 12.5. The left panel presents a genealogy for a sample of individuals, where we make the simplifying assumption that each individual has just one parent in each generation—making things more realistic by having two parents each generation doesn't change any of the overall conclusions but does make it a lot harder to draw (if having only one parent bothers you, imagine that we are tracing the history of mtDNA types from mother to child, or of Y chromosome types from father to son). Note that as we start in the present and go back in time, in each generation the number of ancestors either stays the same or decreases (which happens when two individuals have the same parent). So, eventually the number of ancestors will decrease to one, which then by definition is the MRCA of our sample of individuals.

We can also start in the past with an ancestral population and go forward in time. Each generation, some members of the population leave descendants and some do not. Eventually, all of the members of the population will be descended from just one of the individuals in this ancestral population. So, whether one goes forward in time or backward in time, the overall result is the same: all of the individuals today are descended from a single common ancestor at some point in the past, and that has to be the case as long as there was a single origin of life on this planet.

So far, we are talking only about individuals and not genes, so let's add some mutations. This has been done in the right panel of Figure 12.5, so now you can see how a gene genealogy tracks the genealogy of individuals. As the genealogy shows (and as you already know), not all of the mutations that occurred in the past end up in the present generation. But the mutations that did survive provide us with a means of reconstructing the gene genealogy and, thereby

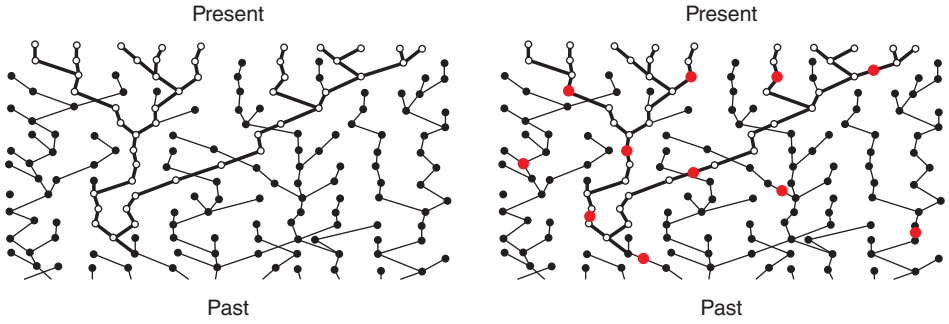


FIGURE 12.5

Relationship between genealogies and mutations. Left, genealogical history of individuals. The thick line and empty circles trace the relationships of individuals alive in the present generation, while the thin lines and solid circles denote the relationships of individuals who do not have descendants in the present generation. Right, mutations (red circles) added to the genealogy.

gaining some insights into the evolutionary history of this population.

These ideas follow from **coalescent theory**, which provides a mathematical framework for constructing gene genealogies and making inferences about population history from them. Coalescent theory is a very useful approach to population genetics, but the math gets rather complex rather quickly and so we won't worry ourselves about the details—those of you who are interested can easily find them elsewhere.

In practice, to estimate the TMRCA for a genomic segment, we need to construct the phylogenetic tree for the DNA sequences (or other appropriate genetic data) using one of the approaches described in the previous chapter, for example, maximum parsimony, maximum likelihood, Bayesian methods, and so forth. If there is no recombination—for example, if we are analyzing mtDNA sequences—then this procedure is relatively straightforward. However, with recombination things get a lot messier, although there are methods for inferring a gene genealogy with recombination. In either event, once we have our estimate of the gene genealogy, then from the average number of mutations that have occurred since the MRCA, and with a molecular clock, we can estimate the TMRCA for that gene. And what does that tell us? Suppose we estimate a TMRCA for a genomic segment in species A of 500,000 years, and a TMRCA for the same segment in species B of 2 million years. What might account for this difference? Before telling you the answer, think about what must be different in terms of the patterns of genetic variation in these two species to give these different TMRCA estimates. Hopefully, you realize that there are more mutations on average from the MRCA to each contemporary sequence in species B than in species A—that has to be the case, by definition, for the TMRCA to be older for species B than for species A (assuming, of course, that the mutation rate is the same in the two species, but you already know how to test that, right?). And how could this come about? A moment's thought should convince you that these differences in TMRCA most likely reflect differences in effective population size: the older the TMRCA, the more mutations there are, and hence the bigger the effective population size. In fact, it can be shown that for an autosomal locus, the TMRCA (measured in generations) is expected to be equal to $4N_e(1 - 1/n)$, where n is the sample size. So, for just two sequences, the TMRCA gives an estimate of $2N_e$, and for reasonably large sample sizes, the TMRCA is about $4N_e$ (since the factor $1 - 1/n$ approaches 1 as n gets bigger). And for mtDNA, the TMRCA provides an estimate of $2N_f(1 - 1/n)$, while for the Y chromosome, the TMRCA estimates $2N_m(1 - 1/n)$, where N_f and N_m are the effective population size for females and males, respectively. Calculating the TMRCA for a genomic segment is thus one way of estimating the effective population size for a species, and we'll discuss this in more detail later in this chapter.

If the effective population size is the only factor influencing TMRCA, then we would expect the TMRCA for all genes to be the same (within statistical error). In fact, other processes—in particular, selection or admixture—can impact TMRCA, and looking for genomic segments with unusual TMRCA estimates is one way of finding candidate genomic segments that have been influenced by selection, admixture, and so forth.

This discussion of TMRCA brings up an important point, namely, that we can measure time either in years or in generations, and for any particular application, you need to be sure to use the right measure. Phylogenetic rate estimates, for example, are usually in years, whereas pedigree rate estimates are typically in generations. The TMRCA estimates are often in years, but if we want to use them to estimate N_e, then the TMRCA has to be in generations. To convert one into the other, you need to know the generation time, and this is not a trivial issue. Generation times for humans that have been used in the literature range from 20 to 35 years, and many studies just assume a generation time without giving the matter much thought. However, a detailed cross-cultural study in 2005 by self-described computer-geek-turned-anthropologist Jack Fenner found longer generation times for males than for females but surprisingly little variation among cultures; estimated generation times for hunter-gatherer societies are quite similar to those for developed nations (Fenner 2005). The recommendation from his study is to use generation times of 25–26 years for mtDNA studies, 30–31 years for Y chromosome studies, and 28–30 years for autosomal DNA studies, and so that is what we will do. Another issue that could complicate the conversion of generations to years is whether or not the generation time has changed in the past, but we won't worry about that for the simple reason that nobody has yet come up with a way to deal with such temporal variation in generation time.

Dating the Age of Mutations

It is straightforward to extend the concept of dating the TMRCA to dating the age of a specific mutation; instead of using all of the genetic variation in a genomic segment of interest, you use only genetic variation that is associated with the mutation of interest. Dating specific mutations sees the widest application in dating mtDNA and NRY haplogroups (which are defined by specific mutations or combinations thereof, as discussed in Chapter 9), so that is what we will focus on in this section. Dating specific autosomal DNA mutations is a messier business because of the

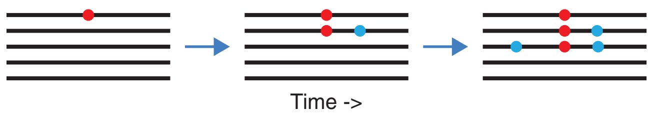


FIGURE 12.6

Diversity associated with a mutation increases over time. A new mutation (red circle) arises, by definition, on a single chromosome (black line), and then over time, new mutations (blue circles) arise and become associated with that mutation.

complications of recombination, so we won't worry about the methods but will mention specific results of interest in later chapters.

The basic idea behind dating specific mtDNA/NRY haplogroups is shown in Figure 12.6. The fundamental concept is that a new mutation arises by definition on a single mtDNA genome or NRY chromosome (let's just call this a mutant chromosome for convenience). Most of the time the new mutation is lost by drift, but occasionally it will rise in frequency and occur as a polymorphism in the population. Initially, all copies of the mutant chromosome will be identical, but over time new mutations will occur at additional sites on the mutant chromosome. So, the genetic diversity associated with the mutant chromosome is correlated with the age of the mutation: the greater the diversity, the older the mutation. Typically, for mtDNA the diversity is based on associated nucleotide substitutions, often only in the hypervariable segments of the control region, but it is becoming increasingly more common to analyze complete mtDNA genome sequences. For an NRY mutation, it is customary to use STR loci to measure the associated diversity, as STR loci evolve quite rapidly, but new approaches based on next-generation sequencing of the NRY promise to shift the focus to nucleotide substitutions for the NRY as well.

By far the most common method used to date specific mtDNA and NRY haplogroups is the so-called "rho" method (after the Greek letter ρ), which is handily implemented in the program Network that is widely used to construct networks of mtDNA sequences and NRY haplotypes (Forster et al. 1996). After constructing the network, all you need to do is click on a few buttons to specify the rate of evolution and the sequences/haplotypes to be included in the dating, and you conveniently get an estimated age and standard deviation. Unfortunately, the ease with which dates can be obtained has meant that insufficient thought often accompanies the analysis. In particular, the rho method is based on the average number of mutations that have occurred among all the sequences descended from a common ancestral sequence and hence assumes a "starlike" phylogeny associated with a population expansion (see Figure 12.7 for an example). In a growing population, by definition each couple is leaving on average more than two offspring (otherwise, the population wouldn't be growing!), so the probability that new mutations will be maintained in the population rather than lost by drift is correspondingly higher than in a population of constant size. Therefore, one expects to see more low-frequency mutations that are derived directly from an ancestral sequence, resulting in a starlike phylogeny (as shown in Figure 12.7). But what if the expansion model is not appropriate and the phylogeny is not starlike? This question was addressed a few years ago by detailed simulations, which showed that whereas with a simulated population expansion the rho method tended to give accurate haplogroup ages, under other scenarios (e.g., constant population size or decrease in size) the dates given by the rho method could be off by quite a bit (Cox 2009). And since there is currently no good way to determine whether the underlying phylogeny is enough like a star to give accurate dates with the rho method, the best advice is to avoid (or at least, supplement) the rho method and use other methods (such as Bayesian or maximum likelihood approaches) that do not depend so heavily on the assumption of a starlike phylogeny.

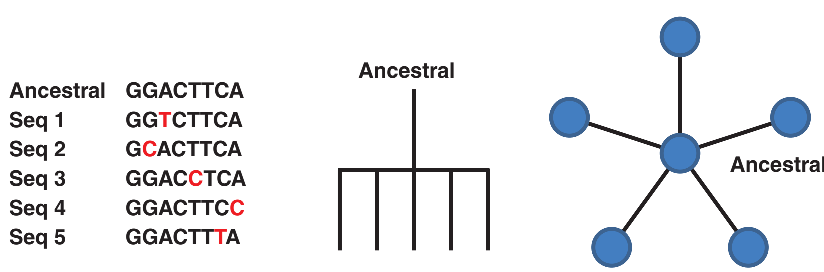


FIGURE 12.7

Schematic illustration as to how a population expansion leads to lots of sequences that are closely related. Left, population expansion results in many sequences that each differ by a single mutation (in red) from the ancestral sequence. Middle, this leads to many branches coalescing at the same time in a phylogeny, which can also be depicted as a "starlike" network (right).

Unfortunately, because it is so easy to obtain haplogroup ages via the rho method, it is still all too common to see studies that use the rho method without any thought as to whether or not it is appropriate for the data at hand. The discriminating reader should be suspicious of any dates obtained solely by the rho method.

What other methods can be used for dating mtDNA or NRY haplogroups (or, more generally, mutations of interest)? Currently, other such methods are based on maximum likelihood or Bayesian approaches. While they have the advantage of flexibility (i.e., they tend to give the correct answer for data simulated under a wide range of population histories), they are computationally intensive and not as easy to implement. Still, with a decent sample of the haplogroup of interest (enough to accurately measure the associated diversity), an appropriate mutational model (i.e., one that incorporates variation in mutation rates among sites for mtDNA, or a stepwise mutation model for STRs on the NRY) to estimate the associated diversity, and an appropriate estimate of the mutation rate, then one can expect to get a reasonably reliable estimate of the age of a haplogroup or mutation of interest (with the inevitable wide confidence limits, of course!).

And how should one interpret an estimate of the age of an mtDNA or NRY haplogroup—what does this actually tell us about population history? Unfortunately, all too often the age of a haplogroup is misinterpreted as corresponding to the age of a migration or a population divergence. In fact, the presence of a haplogroup of a particular age in a population provides only an upper bound as to when that haplogroup entered that population. A haplogroup that arose 50,000 years ago need not have been present in the population for 50,000 years; it could have been contributed to that population by a migration that occurred yesterday. If this is not clear to you, consider the following analogy, which comes from a talk I heard several years ago by the population geneticist Guido Barbujani. Suppose a large population of humans migrate to Mars and some time thereafter are discovered by a Martian biologist. The Martian biologist, intrigued by this interesting new species, carries out an analysis of their mtDNA types. Assuming the humans on Mars are a random sample of those on Earth, then the Martian biologist will find most of the various haplogroups that one finds in humans today (cf. Figure 9.4), with ages ranging from a few thousand to a few tens of thousands of years. But clearly that does not mean that the humans got to Mars tens of thousands of years ago. And yet, it is not difficult to find papers in the literature that report mtDNA or NRY haplogroups that are of various ages and equate each age class with a different migration or event (some even equate each individual haplogroup with a different migration/event, even though it is highly unlikely that a migration would consist of people with only a single haplogroup!).

Those of you who know something about **phylogeography** (which basically consists of using the phylogenetic relationships and ages of haplogroups and their geographic distribution to make inferences about population history) may think that I am denigrating phylogeographic analyses. Nothing could be further from the truth; phylogegraphic analyses provide an extremely informative description of patterns of mtDNA/NRY variation, are an important way to infer the geographic origin of haplogroups (discussed in the next section), and can suggest likely hypotheses as to how the observed patterns arose. However, one of the main themes of this chapter is that it is important to distinguish between the observed patterns of variation and the underlying historical processes that produced the observed patterns. Phylogeography directly informs you about the former but not the latter; to learn about population history requires other approaches that will be described later in this chapter.

Geographic Origin of Mutations

In addition to knowing the age of particular mutations or haplogroups, it is also of interest to know where the mutation/haplogroup arose or from where it started spreading. The basic idea is that a new mutation arises in a single individual at a single location, and as the new mutation gradually increases in frequency over time, it also spreads geographically as people with the mutation move to new locations and reproduce. This suggests that the frequency of a mutation might be one way to infer where it arose: all other things being equal, the frequency of an allele is correlated with its age; the higher the frequency, the older the mutation. So, we might think that where we see the highest frequency of a mutation/haplogroup is where it arose. The problem with this line of thinking is that all other things are seldom equal, and there are many factors besides age that can influence the frequency of a mutation/haplogroup. For example, a severe decrease in population size due to a bottleneck or founder event can result in a large change in allele frequencies. As we shall see in Chapter 16, Polynesians have a very high frequency (95–100%) of a particular mtDNA haplogroup that is also found at lower frequencies in Near Oceania (New Guinea and nearby islands) and eastern Indonesia. However, if you therefore infer that this mtDNA haplogroup arose in Polynesia and spread via a migration westwards from Polynesia, you would be very mistaken. Instead, as we shall see later, other evidence strongly indicates that this mtDNA haplogroup arose in either eastern Indonesia

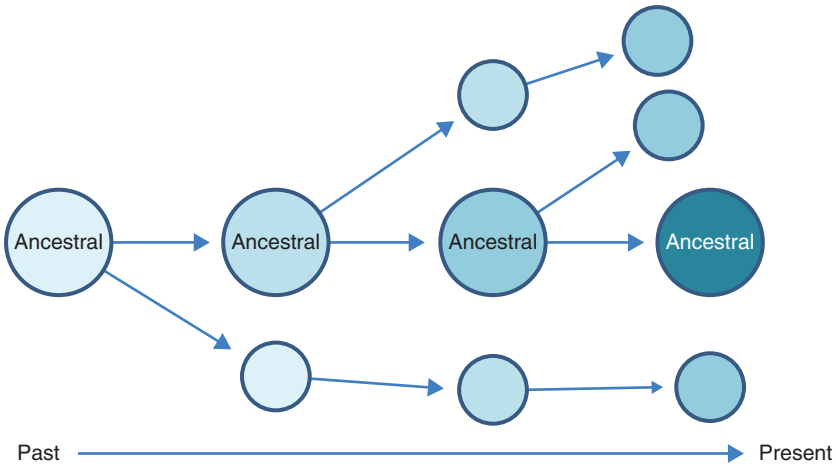


FIGURE 12.8

Schematic depiction as to how genetic diversity associated with a new mutation is expected to be highest in the ancestral population. Circles denote populations; the darker the shading, the higher the diversity. Over time, the ancestral population accumulates diversity from additional mutations, as well as gives rise to new daughter populations that start out with reduced diversity (due to founder events or bottlenecks) and then accumulate additional diversity over time.

or Near Oceania and spread from there eastward to Polynesia; the higher frequency of this haplogroup in Polynesians is explained by the strong founder events and bottlenecks associated with the colonization of Polynesia.

A more informative measure of the origin of a mutation/haplogroup is the amount of genetic diversity associated with that mutation/haplogroup in different populations. The idea again is that a new mutation at a particular nucleotide position arises by definition on a single haplotype, so initially there is no genetic diversity associated with that mutation. Over time, more new mutations occur at other nucleotide positions on the background of that haplotype, so we might expect that where we see the most genetic diversity associated with the mutation/haplogroup of interest, that is the likely origin of that mutation/haplogroup (Figure 12.8). In general, this principle works well, certainly better than simply assuming that the location with the highest frequency is likely to be the origin. For example, even though the frequency of the mtDNA haplogroup characteristic of Polynesians is lower in Near Oceania than in Polynesia, the genetic diversity associated with this haplogroup is considerably higher in Near Oceania than in Polynesia, suggesting a Near Oceanian origin of this haplogroup.

This principle of associating the origin with the highest genetic diversity is based on an underlying model known as the **serial bottleneck** model of migration (Figure 12.9). Basically, if migrating populations are much smaller than the populations from which they originate, then they are expected to carry only a fraction of the genetic diversity of the source population. If the migrant population now settles someplace new, grows somewhat in size, and then becomes the source population for yet another migration, then each successive migration should carry less and less

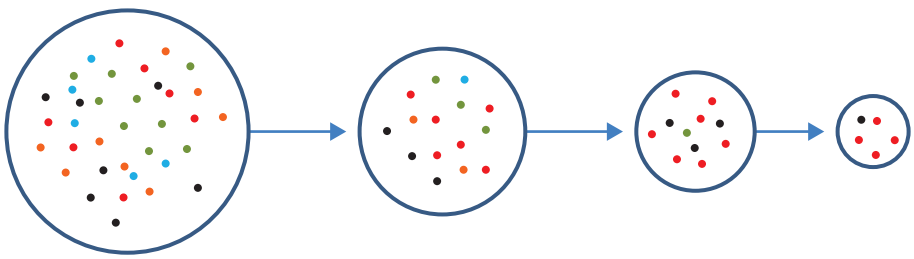


FIGURE 12.9

Depiction of the serial bottleneck model. The founder population (left) gives rise to a smaller population that carries less genetic diversity (colored dots indicate different alleles). This population then gives rise to another population with a smaller size and hence less diversity, and so forth.

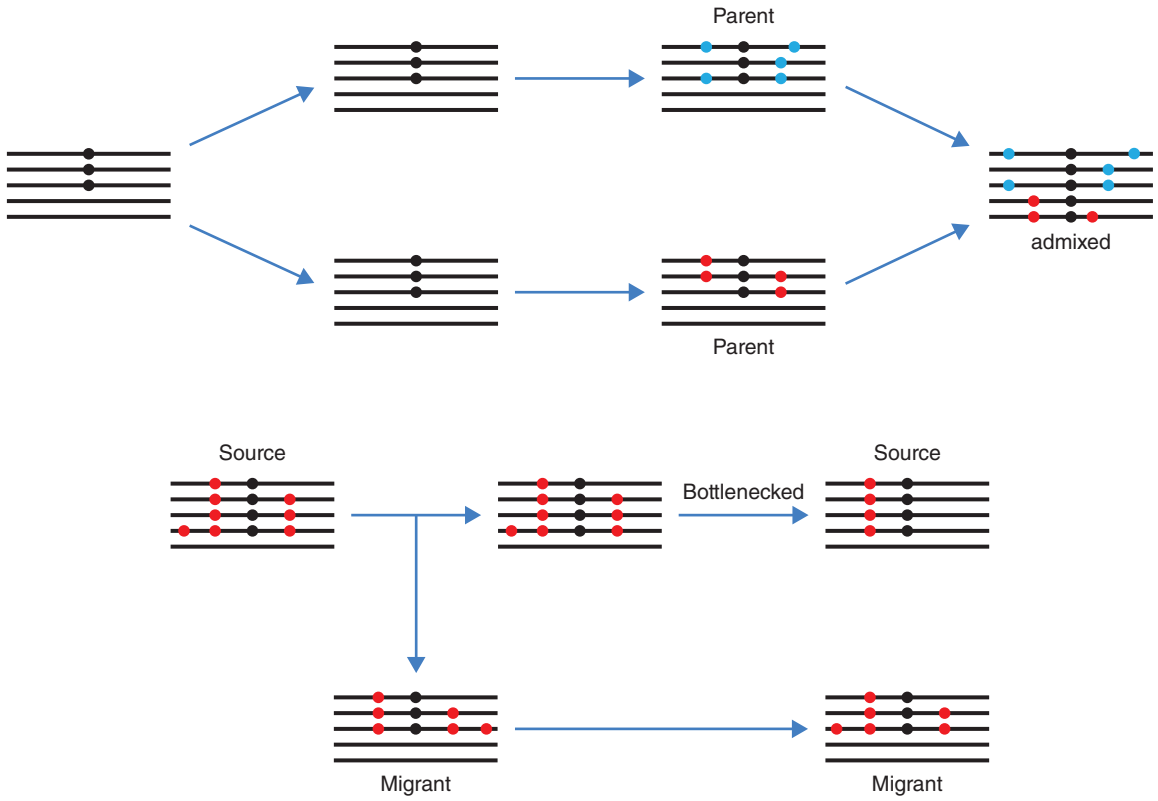


FIGURE 12.10

Two examples where the simple expectation that genetic diversity associated with a mutation should be highest in the ancestral population gives the wrong answer. Lines are chromosomes, black dots are the mutation of interest, and colored dots are additional mutations. Top, admixture: a source population diverges into two daughter populations, each with the mutation of interest, and additional mutations in blue or red accumulate independently. If these two populations then admix at a later time, the diversity associated with the mutation will be higher in the admixed population than in the ancestral population. Bottom, a bottleneck in the source population after a migrant population has left can result in the ancestral population having lower diversity than the migrant population.

of the original genetic diversity. To the extent that this model holds (and, as we shall see, for many human migrations it seems to be a good model), we can infer that the originating population is the one with the highest genetic diversity. But an important theme of this chapter is that while analyses based on models are extremely useful for the insights they can provide, the models always come with underlying assumptions. If the assumptions don't hold, then the inferences may not be valid. For example, suppose a population diverges into two daughter populations that remain isolated for a period of time but then admix; the diversity associated with the admixed population may be higher than in either of the parental populations (Figure 12.10) and could, therefore, lead us to the false inference that a particular mutation/haplogroup arose in the admixed population (African Americans, for example, tend to have more genetic diversity than either Europeans or west Africans). Or, suppose that after a migrating population leaves a source population, the source population undergoes a severe bottleneck, reducing genetic diversity to levels below that of the migrating population (Figure 12.10). These are situations in which the serial bottleneck model does not hold, so we should not be surprised that genetic diversity levels give us false inferences about origins.

Phylogeographic methods offer another way to try to infer the origin of a mutation/haplogroup. The idea here is to use phylogenetic methods to reconstruct the ancestral sequence/haplotype and then look at the geographic distribution of the ancestral type to try to infer where it arose. Typically, where the ancestral type has the highest frequency is taken to be the likeliest origin of the mutation/haplogroup. Phylogeography works best for relatively recent mutations/haplogroups, where the ancestral type is still present at appreciable frequencies in one or more populations. For older mutations/haplogroups, the distribution of the ancestral type in contemporary populations may

be too sporadic (indicating pronounced drift effects) for phylogeography to be of much use.

To summarize, the frequency, associated genetic diversity, and phylogeography of a mutation/haplogroup of interest can all be used to infer where it arose. How much confidence one should have in the resulting inference depends on how well the underlying assumptions of each method are met. A further complication arises from using the location of contemporary populations as a proxy for where their ancestors were in the past; if our analyses indicate that the mutation/haplogroup probably arose in a population that occupies a particular location today, but unbeknownst to us the ancestors of that population were in a completely different place, then our inference will be off. For example, Native Americans carry many mutations that arose in their Asian ancestors; trying to infer the geographic origin for these mutations by studying only Native American populations will not be very productive. Fortunately, the growing use of ancient DNA analyses, discussed in Chapter 15, offers one way of assessing the reliability of inferences about where events happened in the past based on the patterns of genetic variation in contemporary populations. And, as we shall see in later chapters, many useful insights have arisen from genetic analyses that try to figure out where events of interest might have happened, from the origin of modern humans to the source of specific migrations.

Estimating Population Divergence Time

So far we have been focusing on analyses of specific mutations/haplogroups/genetic ancestors (e.g., TMRCA of a particular gene, or when and where a particular haplogroup arose). But since we are most interested in population history, usually we want to know when populations diverged—that is, at what point in time did one common ancestral population diverge to lead to two different populations living today? This is similar to, but not quite the same as, the species divergence question discussed previously. When species diverge, it is assumed that they are no longer capable of exchanging genes, whereas populations can still exchange genes, and such subsequent exchange can have a big impact on estimates of divergence time. In what follows we will assume that once the populations diverge, there was no longer any gene flow or migration between them; in a later section in this chapter, we will consider what happens when we add migration to the mix.

The basic idea behind using genetic data to estimate the population divergence time is that at the time of divergence the two daughter populations are genetically identical (within the limits of random sampling of alleles, of course). Over time, genetic differences will accumulate; the more time since the divergence, the bigger the genetic differences between the populations. There are two sources of subsequent genetic differentiation between two populations, after they diverge from an ancestral population, to consider: (1) alleles that were present in both populations at the time of divergence subsequently changed in frequency due to genetic drift; and (2) new mutations that occurred in either population after they diverged (discussed later in this section). For both of these cases, in principle we can use a molecular clock approach: the amount of genetic difference between the populations, along with a calibration of the rate of genetic differentiation, gives us an estimate as to how long ago the populations diverged.

In practice, in the case of using allele frequency changes to estimate divergence time, it ain't so simple. Estimates of the genetic difference between populations based on allele frequency differences (such as F_{ST}), are influenced not only by time but also by population size changes. Recall from Chapter 10 that under the assumption of no migration, F_{ST} is a function of both time since divergence and the effective population size (N_e). So if we know N_e and F_{ST}, we can estimate the divergence time, and this is an approach that many studies have taken. The main problem with this approach is that while F_{ST} can be estimated precisely from sufficient genetic data, a good estimate of N_e is more difficult to come by (as discussed later in this chapter), so there is great uncertainty in the resulting estimate of divergence time. Even more problematic is the assumption that N_e has been the same in the two populations since they diverged. But if one population grows in size, the rate of allele frequency change via drift will slow down; if a population decreases in size, then the rate of allele frequency change via drift will increase. So, any changes in population size after divergence will further throw off the estimate of divergence time.

Currently, the most useful approaches to dating population divergence times for DNA sequence or genome-wide SNP data are based on using maximum likelihood or Bayesian approaches to first obtain the trees with the highest likelihood or posterior probabilities and then use MCMC (Monte Carlo Markov Chain, described in the previous chapter) or other methods to obtain the best-fitting estimates of population divergence times. There are several such methods available that use different aspects of the data, make different assumptions, and attempt to deal with the problem of ascertainment bias (in the case of genome-wide SNP data obtained from SNP chips). A detailed discussion of these is both beyond the scope of this book and also rather pointless, given that this is an area of very active research and, therefore, current methods are likely to be superseded in the near future by new methods.

We'll therefore reserve any further discussion of specific methods for dating population divergence times for specific applications and examples, as they come up in later chapters.

As just one example of what the new methods might entail, a promising new approach to dating population divergence times is based on identifying new mutations that have arisen in the daughter populations since they diverged (Pickrell et al. 2012). These new mutations are, by definition, polymorphisms found in one population but absent in the other. If we can estimate how many new mutations have arisen, then the amount of time corresponding to these new mutations (which we can get if we know the mutation rate) is our estimate of the population divergence time. This looks to be a very useful approach but does need further evaluation—distinguishing between truly new mutations and older mutations that were either by chance not sampled in one population or were present but lost by drift is not a trivial task. And, we do need an accurate estimate of the mutation rate, which (as will be discussed in the next chapter) is currently a matter of much discussion and debate.

Another potentially interesting approach is based on correlations in patterns of LD (linkage disequilibrium, which was explained in Chapter 9 and involves the nonrandom association of alleles at nucleotide positions that are located close to one another on the same chromosome). The idea here is that at the time of population divergence, the LD between any pair of markers should be the same in the two daughter populations and hence will be perfectly correlated. Over time, since recombination events will be different in the two daughter populations, this correlation will decrease, and this property can be used to estimate the divergence time (McEvoy et al. 2011). However, LD-based divergence times typically underestimate the true divergence time due to fixation of alleles; since in order to calculate the LD between a pair of markers they must both be polymorphic, if an allele at either marker becomes fixed via drift, then that marker drops out of the LD calculation. So, over time, the LD-based divergence time will be based on a biased subset of markers that are still polymorphic, which results in an underestimate of the divergence time.

Perhaps the biggest issue, though, in estimating population divergence time is the often-questionable assumption that, following divergence, the two daughter populations have remained completely isolated, with no further exchange of genes. If there has been subsequent gene flow between the two daughter populations, then this would have the effect of making them appear more genetically similar than in the absence of such gene flow. Thus, the resulting estimate of population divergence time from genetic data would be more recent than the actual population divergence time—and how much more recent is hard to say. For now, unless one has strong evidence from independent sources that the chance of any subsequent gene flow is negligible, the prudent course is to treat any estimate of population divergence time from genetic data as a lower limit to the actual population divergence time.

Some of you, after reading about all of the issues associated with dating population divergence times, may conclude that any such attempt is an exercise in futility. My own view is that, in general, it is better to use a potentially questionable method that may give us imprecise dates than it is to simply throw up our hands and say we can't do it—as long as we are fully aware of the limitations and assumptions of the methods and are careful to state all of the caveats. In other words, a little bit of knowledge is better than none at all, and as we shall see in later chapters, even if we can only say with some degree of confidence that two populations diverged a few thousand years ago versus a few tens of thousands of years ago, that information alone is often enough to tell us something useful.

POPULATION SIZE AND POPULATION SIZE CHANGE

Estimating the effective size of a population is often an important part of population genetics. As we have seen in previous chapters, N_e enters into all sorts of equations and is a necessary component of many methods for estimating population divergence time, among other items of interest. The fundamental idea behind using genetic data to estimate N_e is that the amount of genetic variation in a population is directly proportional to N_e: the more genetic variation, the bigger the N_e. One common way of estimating N_e is from the TMRCA, as discussed previously in this chapter. Other summary statistics concerning genetic diversity that can be derived from DNA sequence data, such as the number of polymorphic sites or average number of nucleotide differences between each pair of sequences, can be used to estimate Θ (introduced back in Chapter 5), which is expected to be $4N_e\mu$ (for autosomal DNA), where μ is the mutation rate. So, with an estimate of the mutation rate, these summary statistics can be used to estimate N_e.

There are also summary statistics based on LD; the idea is that LD is broken down by recombination over time, so the more recombination, the faster the decrease in LD. And since the number of recombination events between two sites is proportional to N_e (because there will be more recombination events in bigger populations), estimates of the amount of LD, along with the recombination rate, can be used to estimate N_e.

The above methods are intended to give a single point estimate of the long-term average N_e for a single

existing population or species. However, we might also want to know N_e for the population that was ancestral to two populations or species—for example, we've already seen back in Chapter 3 that N_e for humans is lower than that for chimpanzees, and that this probably reflects something unusual (such as a bottleneck) during human evolution. But in principle, it could also be the case that a low N_e was characteristic throughout the evolutionary history of our lineage, even before we diverged from chimpanzees, and maybe it's the chimpanzee lineage that is unusual for having a large N_e. Having an estimate of N_e for the population ancestral to humans and chimpanzees would tell us whether N_e for humans decreased or N_e for chimpanzees increased (or both).

One widely used approach to estimate N_e for the ancestral human–chimpanzee population is to add in the gorilla and then examine the discordance between gene genealogies and the phylogeny for these three species. As discussed previously in this chapter (refer back to Figure 12.4), according to the accepted species phylogeny for humans, chimpanzees, and gorillas, we would expect the genome sequences of humans and chimpanzees to be more closely related to each other than either is to the gorilla. And while this is true for most of the genome, for a rather large fraction (about 30%) either the human or the chimpanzee sequence is more closely related to the gorilla. Recall that these discrepancies between the gene genealogy and the species tree reflect incomplete lineage sorting due to polymorphism in the ancestral population (cf. Figure 12.4). The actual amount of incomplete lineage sorting is related to the N_e for the human–chimpanzee ancestor; the bigger the N_e, the more discordance expected between the species phylogeny and the gene genealogies, because the bigger the N_e, the more polymorphism in the ancestral population that traces back to the gorilla. It's not quite so simple, though, because the interval between divergence times also plays a role: the shorter the interval between when the gorilla diverged and when humans and chimpanzees diverged, the more polymorphism you expect to find in the human–chimpanzee ancestral population that is shared with the gorilla. Despite this complication, various methods based on this approach have been developed and give estimates for the N_e of the human–chimpanzee ancestral population of about 50,000–100,000. Since the N_e for humans is estimated to be around 10,000, it does appear as if humans did go through a reduction in N_e after diverging from the common ancestor with chimpanzees. Chimpanzees, on the contrary, have an N_e that is pretty close to the ancestral N_e (around 30,000–50,000), so despite there currently being many fewer chimpanzees than humans, their genetic diversity suggests that for most of their history there were many more chimps than humans around.

All of the aforementioned methods are designed to give us a single point estimate of N_e at a specific time. While such point estimates certainly have their uses, they don't really tell us as much as we would like to know about population history. For example, the best explanation for the relatively low N_e of humans is that at some point in the past we went through a bottleneck—a big decrease in population size—and so there were a lot fewer of us than there are now. Humans may have even been on the verge of extinction, but then clearly the population size not only recovered but also went on to increase enormously to the 7 billion or so humans on the planet today. Obviously, it would be interesting to know when these population size changes occurred, as well as any other changes in population size that are obscured by knowing just the single long-term average value for N_e for humans.

One of the first indications from genetic data concerning past changes in population size came from a type of analysis of mtDNA sequence data called the **mismatch distribution**. Developed by the anthropologists Alan Rogers and Henry Harpending for the analysis of nonrecombining mtDNA sequences (Rogers and Harpending 1992), the mismatch distribution involves taking each pair of mtDNA sequences from a population sample, counting the number of nucleotide differences, and then making a histogram of the number of pairwise differences across the entire sample (Figure 12.11). It turns out that it is pretty simple to figure out what the mismatch distribution should look like for a population that has been constant in size over time: the peak should occur at 0 differences (i.e., most sequences should be identical) and then decrease smoothly (Figure 12.12). With a population expansion, the mismatch distribution will instead have a peak at some nonzero value, with older population expansions having peaks at higher pairwise difference values (Figure 12.12); these are often referred to as waves in the mismatch distribution. A decrease in population size also introduces a wave in the mismatch distribution, albeit with a somewhat different shape that is difficult to distinguish from a population expansion, especially if the population recovers and grows in size after the bottleneck (which should not be surprising, as population growth after a bottleneck essentially is a population expansion).

Most importantly, when there is a wave in the mismatch distribution, the peak of the wave can be used to estimate the time of the population expansion. As an example, Figure 12.13 shows the mismatch distribution for one of the first comprehensive analyses of mtDNA variation in humans. As you can see, there is a pronounced wave in the mismatch distribution, with a peak that suggests a global population expansion in humans at around 50,000 years ago.

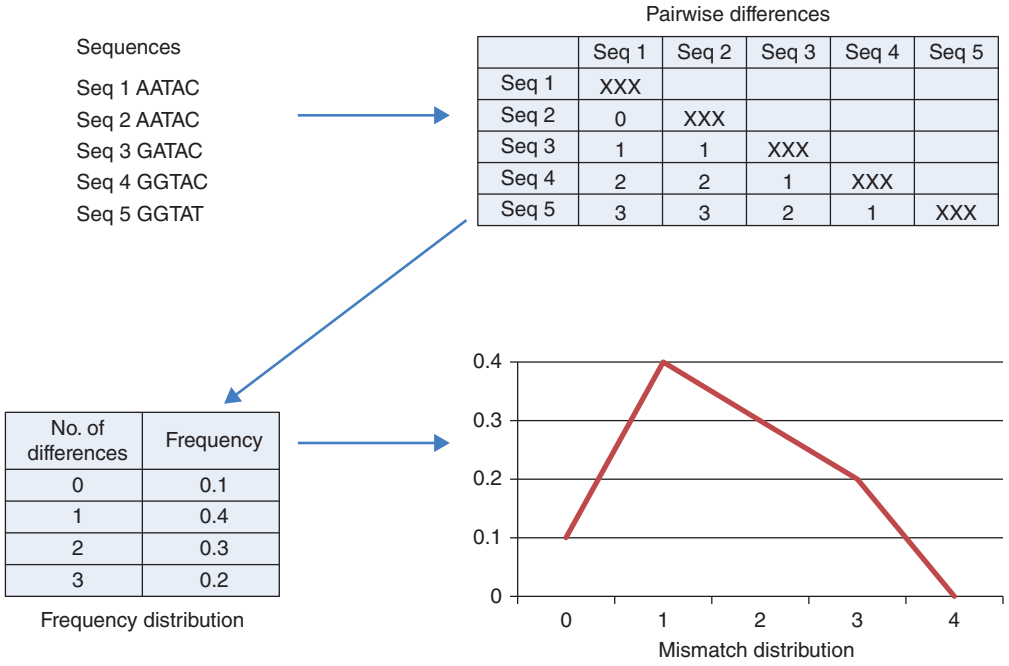


FIGURE 12.11

Schematic illustration as to how a mismatch distribution is obtained. For a set of sequences (top left), the number of mutational differences between each pair of sequences is counted (top right). This is converted into a frequency distribution, which is the observed frequency of pairs of sequences with a given number of mutational differences or mismatches (bottom left), which is then plotted (bottom right) to obtain the mismatch distribution.

Waves in the mismatch distribution are not the only indication of population size changes; the shape of the phylogenetic tree for a sample of sequences also contains information about population size. In particular, a "starlike" pattern in a phylogeny (i.e., a branch with many lineages diverging simultaneously) is also a signature of a population expansion. Figure 12.14 illustrates, for both a constant population and a population that has expanded in the past, the relationship between the distribution of polymorphic sites, the mismatch distribution, and the phylogeny. This

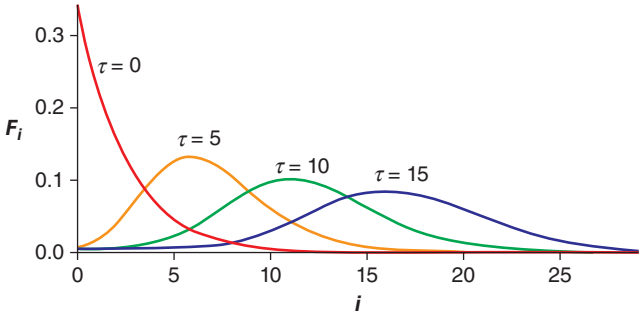


FIGURE 12.12

Expected mismatch distribution after population expansion at different times (τ) in the past. F_i is the frequency of pairs of sequences with i mismatches. With no expansion, the peak of the mismatch distribution is for $i = 0$ mismatches; population expansions create "waves" that have peaks at progressively higher values of i for correspondingly older population expansions. Modified with permission from Rogers, A., and Harpending, H., "Population growth makes waves in the distribution of pairwise genetic differences," *Molecular Biology and Evolution* 9:552, 1992.

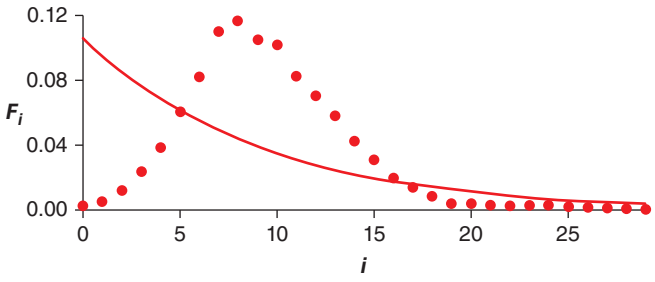


FIGURE 12.13

The mismatch distribution expected under neutrality (solid line) compared to that observed in human mtDNA data (solid circles). F_i is the frequency of pairs of sequences with i mismatches. Modified with permission from Rogers, A., and Harpending, H., "Population growth makes waves in the distribution of pairwise genetic differences," *Molecular Biology and Evolution* 9:552, 1992.

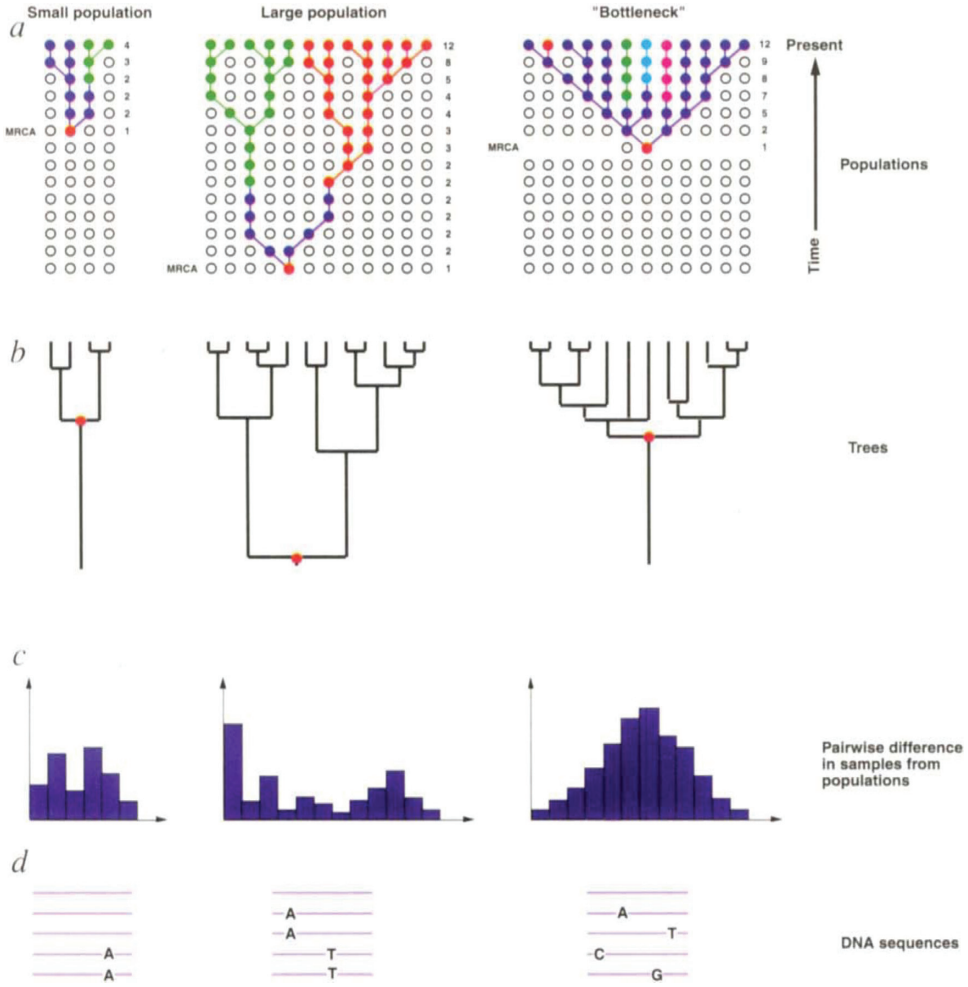


FIGURE 12.14

Correspondence between (a) population history, (b) trees, (c) mismatch distributions, and (d) DNA sequence variation for different population sizes. Reprinted with permission from von Haeseler, A., et al., "The genetical archaeology of the human genome," *Nature Genetics* 14:135, 1996.

relationship suggests that all three can provide information on past population size changes; having discussed mismatch distributions, we'll now turn to the analyses of polymorphic sites and of the phylogeny.

The usual way to analyze data on polymorphic sites, arising either from sequencing or from SNP genotyping, is based on the **allele frequency spectrum** (or AFS). The AFS is typically based on polymorphic sites that have two alleles and can be either "unfolded" or "folded"; the unfolded AFS assumes that we know which allele is ancestral and which is derived at each site (usually, by comparison to one or more outgroups), while the folded AFS assumes that this information is not known. For the unfolded AFS and a sample of n chromosomes (so, $n/2$ individuals), we count the number of sites with one derived allele and $n - 1$ ancestral alleles, two derived alleles and $n - 2$ ancestral alleles, and so forth up to $n - 1$ derived alleles and one ancestral allele (if the count of the derived allele is either 0 or n, then the site is not polymorphic). The result is a histogram of the frequency of each possible derived allele count in the sample, as shown in Figure 12.15. For the folded AFS, since we don't know which allele is derived and which is ancestral, we focus on the minor allele (i.e., the one present in at most $n/2$ chromosomes) and count the number of sites with one minor allele and $n - 1$ major alleles, two minor alleles and $n - 2$ major alleles, and so forth, up to a maximum of $n/2$ minor alleles and $n/2$ major alleles; the result is a histogram of the frequency of each possible minor allele count in the sample (also shown in Figure 12.15).

It turns out that the AFS captures a good deal of the information in the data, and the expected AFS can be easily calculated for a constant-size population. Moreover, the AFS varies in characteristic ways according to how the population size has changed in the past (Figure 12.16). Population growth results in an excess of

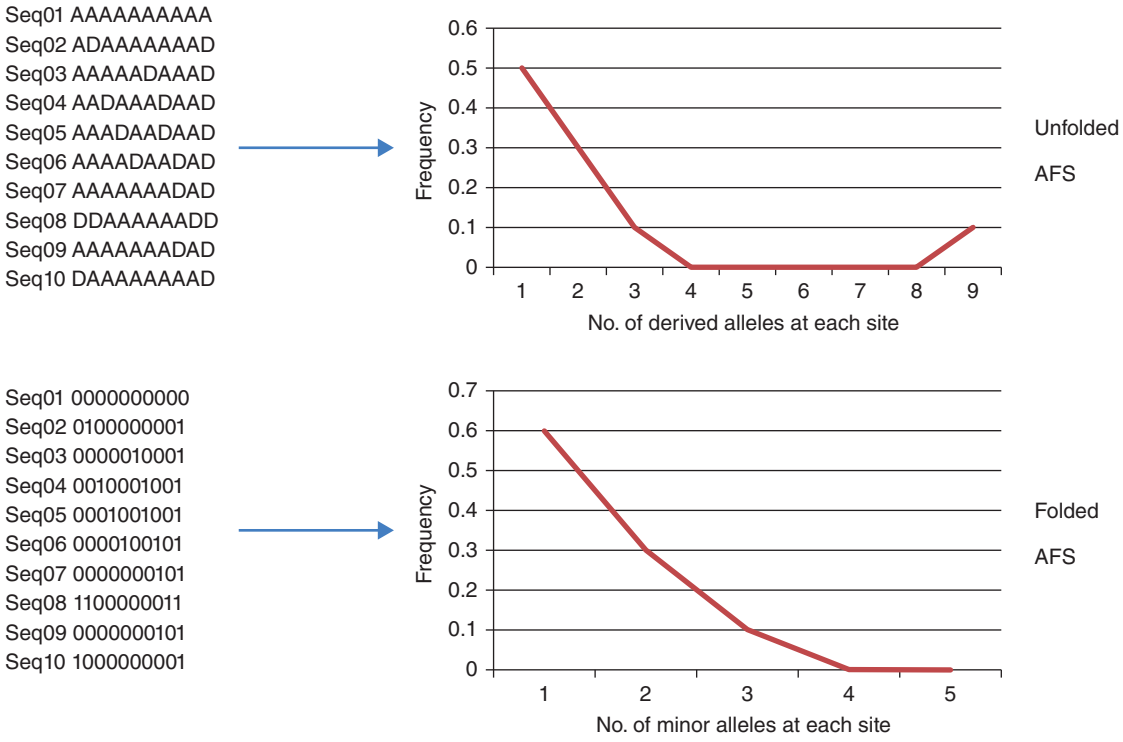


FIGURE 12.15

Folded and unfolded AFS. Suppose we have 10 sequences (top left), with A indicating the ancestral allele and D the derived allele. The number of derived alleles at each polymorphic site, therefore, can range from 1 to 9, and the unfolded AFS is the frequency of each count. If we don't know which allele is ancestral and which is derived, then (bottom left) we have polymorphic sites (with the alleles designated 1 and 0) and we count the number of minor (less frequent) alleles (which can range from 1 to 5) to get the folded AFS. AFS indicates allele frequency spectra.

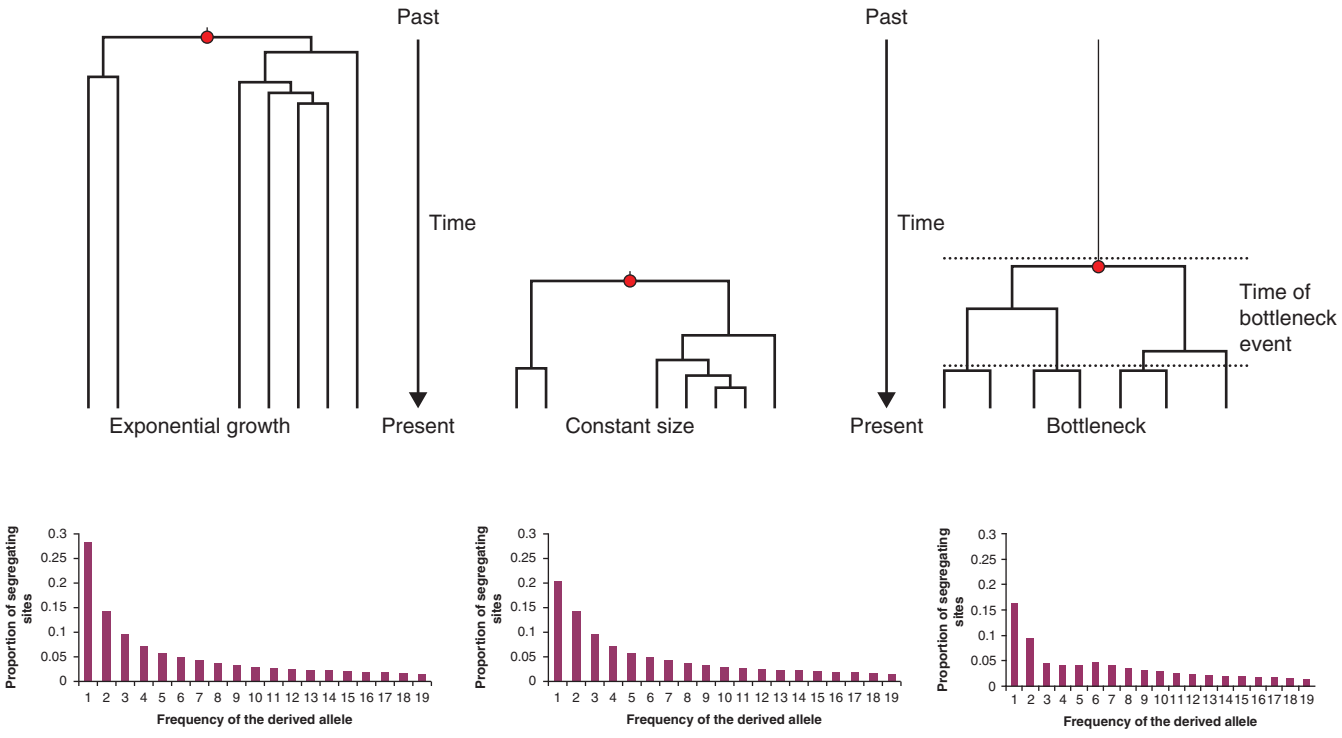


FIGURE 12.16

Expected AFS for different demographic scenarios. AFS indicates allele frequency spectra.

low-frequency alleles (refer back to Figure 12.14 if this is not obvious to you), while a decrease in population size results in a deficiency of low-frequency alleles and an excess of intermediate-frequency alleles, because the low-frequency alleles are lost via drift more often when the population size is small.

The usual approach to estimating population size change in the past from the AFS is to simulate data under various models of population size change and see which model gives the best fit to the data. The typical sorts of scenarios are illustrated in Figure 12.17, which shows a population at a certain constant N_e (N_0) that then undergoes a sudden decrease in size at time T_1 generations in the past to N_b, which is then followed by exponential growth to the present N_z. Note that this is a flexible scheme, as it can also accommodate population growth without any prior reduction in size (in which case, $N_b = N_0$) as well as a population decrease without any subsequent increase in population size (in which case, $N_z = N_b$). Various approaches have been suggested as to how to obtain estimates of the desired demographic parameters from the AFS, and this is a very active area of ongoing investigation.

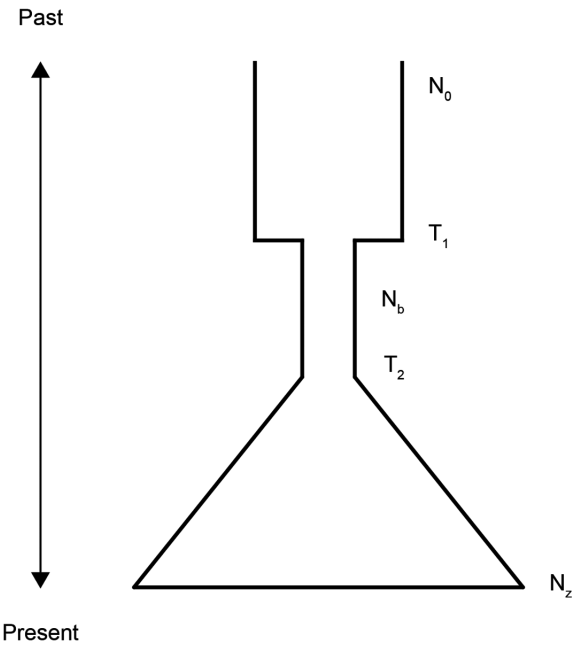


FIGURE 12.17

Generalized model of population size change from past to present. A population begins with an effective size of N_0, goes through a reduction in size at time T_1 to an effective size of N_b, then starts expanding at time T_2 to reach the present day effective size of N_Z. Although the figure shows specifically a population decrease followed by expansion, in practice any population size change of any duration at any time can be modeled.

While the aforementioned modeling approach can be quite informative, it does have the drawback of trying to fit the data to a specific model that (usually) includes just one episode of population size change. Real population history is undoubtedly more complex, with potentially multiple episodes of both population growth and reduction. One type of analysis that tries to infer the entire history of population size change produces what are called **Bayesian skyline plots** (or BSPs for short) (Drummond et al. 2005). The principle behind BSPs is illustrated in Figure 12.18; the basic idea is that if you have a phylogenetic tree that is consistent with a molecular clock, then you can date each branching event in the tree. Moreover, the number of branching events in a given time interval is going to be proportional to the N_e for the population during that time: the more branching events observed, then the bigger the N_e was during that time. To construct a BSP, you use Bayesian methods to estimate the tree, the dates for the branching events in the tree, and the population size for each time interval in the tree. The end result is a plot of population size change over time—the BSP, along with associated confidence limits for the population size (which

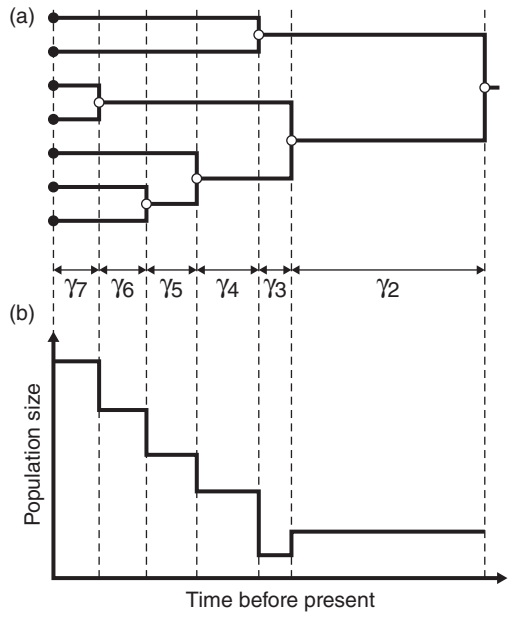


FIGURE 12.18

Rationale behind Bayesian skyline plots. With a tree that relates sequences with time points assigned to branching events (a), corresponding estimates of the effective population size (b) that are most consistent with the history of branching events in the tree can be obtained. Reprinted with permission from Ho, S., and Shapiro, B., "Skyline-plot methods for estimating demographic history from nucleotide sequences," *Molecular Ecology Resources* 11:423, 2011.

usually are quite considerable). Note that BSPs are usually constructed for mtDNA sequences, since it is (relatively) straightforward to produce the tree and associated dates of branching events for nonrecombining mtDNA sequences. And, with increasing numbers of partial Y chromosome sequences generated via next-generation sequencing, we are starting to see BSP plots based on Y chromosome sequences (e.g., Lippold et al. 2014). However, with the advent of genome-wide data sets, there is considerable interest in analyzing population size change over time from such data, which in principle should provide considerably more information than can be gleaned from a single genetic locus such as mtDNA.

Among recent attempts to provide BSP-like plots from genome-wide data is a remarkable method that uses the genome sequence from a single individual to infer the history of size changes for the population from which that individual is derived. You may wonder how one can possibly infer population size change over time from genetic information from a single individual. The answer is that our genomes consist of millions of independent loci, consisting of haplotype blocks (genomic segments that are inherited as a unit) that are bounded by recombination events. Each independent locus has an independent history, and thus our genome sequence is actually a population of genomic segments, each with their own history. Moreover, with diploid sequences, the TMRCA (time to most recent common ancestry) can be estimated for the two sequences at each locus, based on the number of heterozygous and homozygous sites. Therefore, in principle, it should be possible to get a distribution of TMRCAs across the genome, and then one can use this information (using the same principle as in Figure 12.18) to estimate the most likely N_e at each time point.

So in theory, one can estimate the history of population size change from a single diploid sequence. While this was well-known, in practice nobody had a good idea as to how to actually do this, until researchers Heng Li and Richard Durbin developed a novel method called the pairwise sequential Markovian coalescent model (Li and Durbin 2011). The details are quite complex, but the basic idea (Figure 12.19) is that you move sequentially along a chromosome until the distribution of homozygous and heterozygous sites becomes too heterogeneous to be modeled by a single history. The inference is that there must then be a recombination event in this region of the chromosome, in order to produce adjacent segments with different histories. One then defines a new segment and repeats the process, thereby dividing each chromosome into homogeneous segments bounded by recombination events. For each homogeneous segment, the TMRCA can be estimated, thereby producing a genome-wide distribution of TRMCAs that can then be used to estimate the most likely history of population size changes that would produce the observed TMRCA distribution. Simple, right? We'll see an example of

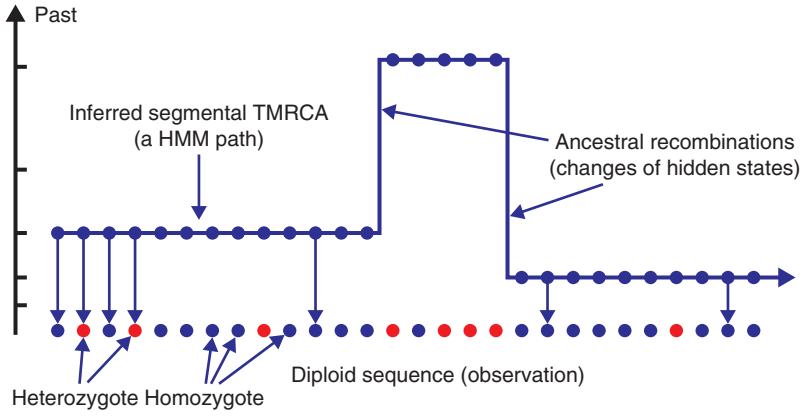


FIGURE 12.19

Logic behind using pairwise sequential Markovian coalescent (PSMC) to infer population size change through time. A chromosomal sequence from a single individual consists of homozygous and heterozygous positions; when the distribution of homozygous and heterozygous sites becomes too heterogeneous to be consistent with a single history (TMRCA) for that chromosomal segment, then a recombination event is inferred. The length of the chromosomal segment provides an estimate of the age of that segment (the shorter the segment, the older it is) while the TMRCA provides an estimate of the effective population size (N_e) for that segment (the older the TMRCA, the bigger the N_e). The output of the PSMC method is thus a large number of segments that can be binned into different ages to give the distribution of N_e at different times in the past (similar to Bayesian skyline plots). TMRCA, time to the most recent common ancestor. Modified with permission from Li, H., and Durbin, R., "Inference of human population history from individual whole-genome sequences," *Nature* 475:493, 2011.

a pairwise sequential Markovian coalescent analysis later on in Chapter 14 (see Figure 14.15). Incidentally, it took more than 2 years from the time this paper was submitted until it was accepted for publication in the prestigious journal *Nature*, which must be some kind of record—2 years is a lifetime in this modern genomics era!

MIGRATION AND ADMIXTURE

If there are two things humans like to do, it's migrate and mate, and the result is known as **admixture**, or the contribution of genes from one population to another. We have already seen some descriptive methods that can indicate that admixture might have occurred, such as the PC and STRUCTURE-like analyses described in the previous chapter. Here, we want to focus on determining which populations were involved in the admixture and then estimating how much ancestry each parental population contributed to the admixed population and when admixture events might have occurred in the past, as indicated in the simple model in Figure 12.20. Although some attempts have been made to apply formal admixture analyses to mtDNA and/or NRY data, any estimates based on single loci inevitably lack precision and hence aren't terribly useful. Still, by using phylogeographic methods one can often get a pretty good idea as to where particular mtDNA or NRY haplogroups arose and thereby make some inferences about admixture that can be particularly useful when the admixture has been sex-biased; we'll see an example in Chapter 16. But for the remainder of this section, we will focus on analysis of admixture from genome-wide data, as that is where we can get enough information to make reasonable estimates of the amount and time of admixture.

A fundamental assumption of most methods for estimating admixture is that the parental populations that participated in the admixture (i.e., the populations that contributed genes to the admixed population), or reasonable proxies thereof, have been correctly identified. Unfortunately, many commonly used programs to estimate admixture won't tell you that there may be a problem with your choice of parental populations but instead will go ahead and do their best to estimate the admixture parameters. If you want to model African Pygmies as the result of admixture between New Guinean Highlanders and Greenland Eskimos, these programs will do their best to give you an answer, no matter how nonsensical.

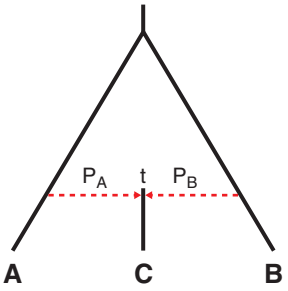


FIGURE 12.20

Simple model of admixture. Populations A and B diverged in the past and then give rise at time *t* to population C, with P_A and P_B the amount of ancestry contributed to C by A and B, respectively.

A related issue is that since admixture by definition happened in the past, but we don't have access to the gene pool of the parental populations at the time of admixture, we have to rely on current populations that we think best represent the parental populations (i.e., proxies). But current populations, even if they are highly likely to be directly descended from the parental populations, may poorly represent the genetic composition of the populations involved in the admixture—genetic drift as well as subsequent admixture events may have considerably altered the gene pool of the proxies. This can be especially problematic for admixture that is inferred to have happened a long time ago. The take-home message: always pay attention to the choice of proxies for the parental populations in any investigation of admixture, especially when reading the older literature, where sometimes this issue was not carefully considered.

Fortunately, we don't have to simply rely on best guesses or descriptive analyses (such as PC or STRUCTURE-like analyses) to guide our investigation of admixture, as recent developments have made it possible to formally test whether or not there is evidence of admixture involving specified populations. Various statistical tests have been developed for this purpose; currently one of the most commonly used such tests is the **f_4 test** (Reich et al. 2009), and the rationale behind this test is depicted in Figure 12.21. The idea is that we start with a tree that depicts what we think is the true history of four populations, assuming no admixture. For each polymorphic position in the data, we calculate the difference in allele frequencies between populations A and B and between populations C and D. We then calculate the correlation in these allele frequency differences (A-B vs. C-D). As the left panel in Figure 12.21 illustrates, if there has been no admixture and the tree is correct, then the A-B allele frequency differences should not be correlated with the C-D allele frequency differences, because they involve completely different parts of the tree; there is no shared history when comparing A-B versus C-D, so, therefore, allele frequency differences that arise via genetic drift in different parts of the tree should not be correlated. The expected value of the f_4 statistic is

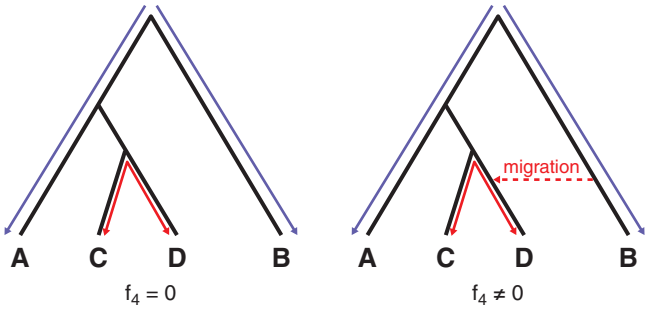


FIGURE 12.21

Rationale behind the f_4 statistic. If the tree on the left accurately describes the history of these populations, then the correlation in allele frequency change between A and B (related by the blue part of the tree) versus C and D (related by the red part of the tree) should be zero, because there is no overlap between these two comparisons in the tree. But if the tree is not accurate because of migration (in this case, from B to D), then the changes in allele frequencies will not be independent in A-B versus C-D, and so the f_4 statistic will be significantly different from zero.

thus 0. However, if there has been admixture from one population into another (as depicted in the right panel in Figure 12.21), then this will introduce a correlation in the allele frequency differences when comparing A-B versus C-D, and the f_4 statistic will be significantly positive or negative (depending on which populations are involved in the admixture). These f_4 statistics can be readily calculated for all populations of interest, thereby providing statistical support for evidence of admixture, and are seeing increasing use in studies with appropriate genome-wide data.

The f_4 tests require population data in the form of allele frequencies; a related test that can provide evidence of admixture in single genome sequences is the so-called **D statistic** (Green et al. 2010). So far, the D statistic has been largely applied to looking for evidence of admixture from archaic hominins (AHs) (such as Neandertals or Denisovans, discussed in Chapter 15), and the rationale behind the D statistic is illustrated in Figure 12.22. As with the f_4 statistic, we start with an assumed phylogeny, this time for three sequences: one from an AH and two from different modern humans (H_1 and H_2). We then focus on biallelic nucleotide positions in which we can infer (by comparison to the chimpanzee and/or other nonhuman primate sequences) which allele is the ancestral state (A) and which is the derived state (D). If we then examine nucleotide positions that differ among the three sequences, most of the time AH has the A allele and H_1 and H_2 both have the D allele; for these positions, the gene genealogy matches the species tree. For the D statistic, we focus instead on positions in which AH has the D allele while one human sequence has the A allele and the other has the D allele. There are two possible explanations for such discrepancies between the gene genealogy and the species tree (Figure 12.22); the first is that this reflects polymorphism in the ancestral population for AH and humans. That is, the mutation producing the D allele occurred prior to

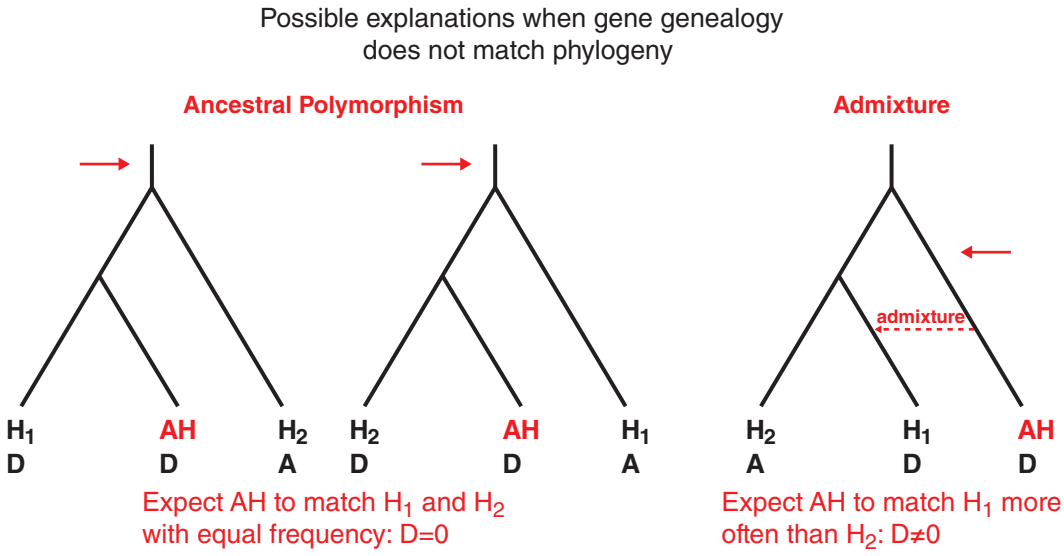


FIGURE 12.22

Use of D statistics to distinguish between two possible explanations for a mismatch between the gene geneaology and the expected relationships for a set of sequences. AH indicates archaic human; H_1 and H_2, modern humans. The arrow indicates mutation from the ancestral (A) to the derived (D) allele. See text for further details.

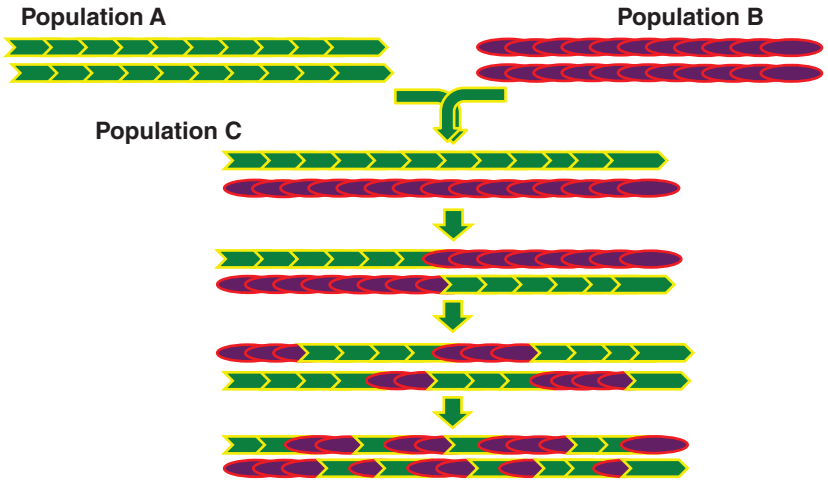


FIGURE 12.23

Evidence used to estimate the time of admixture. When populations with two different ancestries (A, B) interbreed to form a third population (C), initially there will be large chromosomal segments of ancestry from each parental population. Over time, the size of the segments is reduced by recombination and they become more numerous—with estimates of the recombination rate, these two properties can be used to estimate the time of admixture events. Reprinted with permission from Pugach, I., et al., "Dating the age of admixture via wavelet transform analysis of genome-wide data," *Genome Biology* 12:R19, 2011.

the divergence between AH and humans, and so both the A and D alleles were present in the common ancestral population of AH and humans. By chance, AH and one human have the D allele and the other human has the A allele. The key point here is that if ancestral polymorphism is indeed the explanation for the D alleles in AH and just one of the two humans, then the chance that H_1 has the D allele is the same as the chance that H_2 has the D allele. Therefore, the number of times AH and H_1 (but not H_2) have D alleles is expected to be equal to the number of times that AH and H_2 (but not H_1) have D alleles, and the D statistic is expected to be equal to 0. The second possible explanation (Figure 12.22) for AH and one of the two humans having the D allele is that there was admixture between AH and the ancestors of that one human; in this case, mutations that occurred on the AH lineage will be shared between AH and the admixed human, and hence there will be an excess of D alleles shared between AH and that human. In this situation, the D statistic is expected to be significantly different from 0, with an excess of D alleles shared between AH and the human descended from the admixed population. We'll see later what this test has to tell us about admixture between our ancestors and AHs.

Once we have convinced ourselves that we have formal evidence of admixture and have identified the appropriate parental populations, then estimating the amount of admixture is rather trivial; we have already seen how to do this for a simple case involving a single locus back in Chapter 5, and this can easily be extended to multiple loci in genome-wide data. Estimates of the amount of admixture can also be obtained from STRUCTURE-like analyses and other methods.

Of more interest is trying to figure out when the admixture occurred, as this can tell us about migration events in the past. The basis for many approaches for using genome-wide data to date admixture is illustrated in Figure 12.23; at the time of admixture, an admixed individual will have one member of each chromosome pair of one ancestry and the other member of the other ancestry. Over time, recombination will break down the blocks of ancestry, so that they become smaller and more numerous. Several methods have been developed to estimate admixture time from the number and/or size of the admixture blocks (or related properties thereof), and these are seeing increasing use. Other approaches take advantage of the fact that admixture is expected to increase LD in the admixed population, and the degree of LD introduced by admixture should then be correlated with the genetic differentiation between the parental populations, and from this correlation the time of admixture can be estimated. New methods are constantly being developed, as this is an area of active and fertile research. Currently, most methods for estimating admixture time assume a one-time admixture event, but progress is already being made on estimating the duration of admixture (when it occurs over several generations) as well as the times of multiple admixture events.

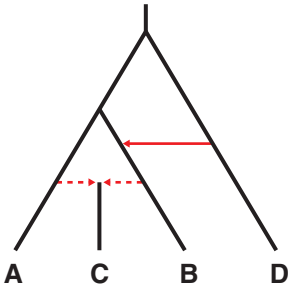


FIGURE 12.24

Schematic depiction as to how a population can have ancestry from another population without direct interactions between the ancestors. In this scenario, ancestors of population D contribute genes to ancestors of population B. Subsequently, some ancestors of population B admix with ancestors of population A to form population C. Population C thus has ancestry from population D, even though the ancestors of C and D never interacted directly.

One important point to make about interpreting admixture signals in genetic data: the presence of ancestry from different populations in an admixed population does not necessarily indicate that the parental groups involved actually met and admixed (as implied, e.g., in Figure 12.20). It could be that the admixed population received the ancestry via other populations (e.g., Figure 12.24). In the scenario depicted in this figure, admixed population C has a signal of ancestry from population D, even though the ancestors of D never interacted directly with the ancestors of C. I raise this point because some geneticists do not make this distinction; to them, whether the scenario shown in Figure 12.20 or in Figure 12.24 depicts the actual history is irrelevant, because in either case population C has ancestry from population D and that is what is important. While in one sense this is true, in another sense the two scenarios in Figures 12.20 and 12.24 have very different historical implications as to how the genes from population D ended up in population C, and if we are interested in explaining the underlying history behind observations of shared genetic ancestry, we have to keep these distinctions in mind.

One last point about admixture: as we get better and better at detecting the signal of admixture in genome-wide data, one of the most important outcomes is that we see more and more evidence of previously unsuspected admixture events, and not just between AHs and modern humans but also between different human populations; we'll see some examples later in this book. Perhaps, given the human proclivity for migrating and mating, this should not be so surprising. Nevertheless, not so long ago it was not so uncommon at meetings discussing human genetic diversity to hear some researchers assert that prior to 1492, most (non-European) human populations lived (more or less) in a state of relative isolation, with "pristine" gene pools, and it was only with the advent of European-driven colonization that populations started to significantly mix genetically. While one could debate how reasonable this view was even back then, there certainly is no basis for it nowadays: when it comes to humans, admixture is the rule, not the exception, and that has been true throughout our evolutionary history, not just recently.

PUTTING IT ALL TOGETHER

So far, we have considered the various aspects of demographic history that we would like to learn about from genetic data separately, for example, population divergence time, changes in population size over time, admixture events, and so forth. But in reality, all of these may have occurred during the history of particular populations (cf. Figure 12.1) and will influence patterns of genetic variation accordingly, so it would be desirable to try to estimate all of the demographic parameters of interest simultaneously. There are several approaches that vary in the details but for which the logic is more or less the same. A commonly used approach is called **approximate Bayesian computation**, or ABC for short (Beaumont et al. 2002). The idea is that you start with a model of population history and a set of parameters that you would like to estimate—take Figure 12.25, for example, in which there are six demographic parameters of interest (ancestral N_e, N_e in each of the two daughter populations, time of divergence, and average migration after divergence from population A to population B and vice versa). You have collected genetic data (sequence data or genome-wide SNPs) from samples from the two daughter populations, so the goal is to estimate these demographic parameters from the genetic data. The idea then is to simulate similar genetic data a large number of times to encompass a wide range of possible parameter values and then choose those simulations that most closely match the observed genetic data and use the corresponding parameter estimates. Since in general it is not computationally feasible to calculate the full likelihood of the data given each combination of possible parameter values, what one does in practice is to calculate **summary statistics** from the observed and simulated data. These summary statistics are chosen to (hopefully) capture as much of the information in the data as possible; commonly used summary statistics would include nucleotide diversity, heterozygosity, number of polymorphic sites, F_{ST}, and so forth. Because the analysis is based on summary

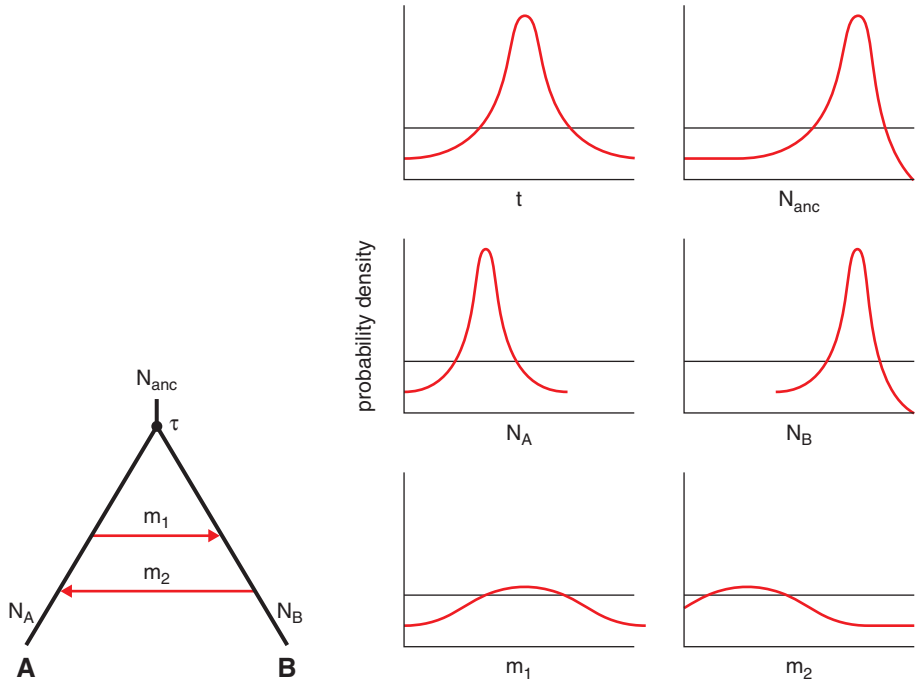


FIGURE 12.25

Typical model that might be used in ABC analysis. The left panel shows the modeled history for two populations, with six parameters of interest: time of divergence (t), effective size of the ancestral population (N_{anc}) and the two existing populations (N_A and N_B), and average migration rate from A to B (m_1) and from B to A (m_2). The six plots show the prior (black) and posterior (red) distributions obtained for the six parameters. Note that the ABC analysis provides relatively precise estimates of the first four parameters but relatively poor estimates of the two migration rates.

statistics and not the full likelihood of the data, it provides an approximate answer that is only as good as the summary statistics (and this is why it is called "approximate" Bayesian computation). In practice, you start with prior distributions for each of your parameters of interest (this is where the Bayesian part comes in), select values for each parameter from the prior distribution, simulate genetic data under this history, calculate the summary statistics for the simulated data, and then compare the simulated summary statistics to the observed values. If these are close enough (usually, you specify a threshold for how different acceptable summary statistics can be), then you keep the parameter values, otherwise you reject them (so, ABC is also sometimes referred to as a rejection algorithm). You then repeat the process until you have accumulated a predefined number of simulations with acceptable summary statistics. The distribution of values for each parameter among these acceptable simulations then provides the posterior distribution, from which the best estimate of the parameter value and approximate confidence intervals (i.e., the range for which 90% of the values fall) can be derived (cf. Figure 12.25). Note that in this hypothetical example relatively good posterior distributions were obtained for the time of divergence and the various N_e values, but the migration rates are poorly estimated (because the posterior distributions are quite close to the prior distributions). A common outcome of such analyses is that some demographic parameters are easier to estimate from genetic data than others, so a nice feature of the ABC approach is that in addition to your parameter estimate, you also get some idea as to how much information in your parameter estimate is coming from the genetic data versus how much is coming from your prior estimate (i.e., if the prior and posterior distributions are quite similar, then the genetic data are providing very little in the way of additional information).

The basic ABC approach described previously can still take a lot of computational time (on the order of weeks or even months, depending on the complexity of the model and the data), so one much-used improvement is to combine ABC with MCMC (Markov chain Monte Carlo); this combined approach has the imaginative name of ABC–MCMC. The problem with the simple ABC approach is that the parameter values for each simulation are chosen at random from the prior distributions, so you may spend a lot of your time running simulations with suboptimal parameter values

that have no chance of ever passing the rejection step. In ABC–MCMC, you start as with ABC by choosing parameter values from the prior distributions, doing the simulations, and calculating summary statistics to compare to the observed summary statistics, until you arrive at a set of parameter values that yield simulated summary statistics that pass your threshold value (or, if you are really clever, you specify beginning parameter values that you think might be close to the real values). You then use MCMC to move from these parameter values to another set of parameter values, do the simulation, calculate the summary statistics, and if these simulated summary statistics also pass your threshold value, you keep them and continue the MCMC. You continue this process until you have the specified number of simulations that pass the specified threshold value when compared to the observed summary statistics. So, the advantage of ABC–MCMC is that more of your simulations should be with acceptable values of the parameters, decreasing the overall number of simulations that need to be done and hence the computational time (of course, the usual caveats with MCMC hold, such as surveying enough different parameter values to be sure you aren't missing the best ones, etc.)

ABC–MCMC isn't the only approach that can be used to infer multiple demographic parameters, but the logic is more or less the same for all of the current approaches (e.g., simulate population history and compare simulated values to observed values), so we won't worry about the details of other approaches. Again, this is an area of very active and intense research, and it is quite likely that in the not so distant future, new approaches will render current ones obsolete.

One final word about models: an important caveat to keep in mind is that the modeled history that provides the best fit to genomic data is not necessarily the true history. Genomic data give us a window into the past, but the past is by definition a very long time and may encompass lots of events that we have no knowledge of that nonetheless have shaped patterns of genetic variation in current populations. If the underlying model used to generate estimates of demographic parameters is incorrect, then obviously all bets are off when it comes to interpreting the parameter values. This is not to disparage the use of genetic data to investigate the past, nor the modeling approach; as we shall see in the next chapters, genetic data have provided some important insights into our evolutionary history that (arguably) would not have arisen otherwise. And models have the advantage that they at least make clear what one is assuming has happened in the past; these assumptions can be tested to see how strongly they impact the analyses (e.g., by seeing what happens if you make different assumptions), and they can be updated as our knowledge of the past improves. One of the big advantages of studying the genetic history of humans (as opposed to other species) is that we have a rich source of information about our past coming from paleontology, archaeology, linguistics, and many other fields; thus, insights from genetics can be compared to insights from these other areas, thereby providing a comprehensive and (hopefully) unified view of our past history.

CHAPTER 13

OUR CLOSEST LIVING RELATIVES

The previous sections of the book have provided background for the uninitiated concerning various aspects of genetics, populations, evolution, and the production and analysis of molecular genetic data. Now, we are at last ready to see what we have learned about human evolution and population history from molecular anthropology. We will begin with what is (arguably) one of the most significant contributions of molecular anthropology: namely, the answer to the question, who are our closest living relatives?

As we delve into the answer, there is an important issue to keep in mind. A continuing problem with studies of human evolution is that the subject of the investigations is the same as the investigators—that is, we are studying ourselves. While science demands objectivity, at the same time (consciously or subconsciously) we tend to think that we are pretty special among all of the various living creatures, and so (consciously or subconsciously) we expect to see evidence in our evolutionary history as to how we became so special. This is not to deny that humans do have some rather special properties; we are, after all, the only species on this planet that writes textbooks about their origins and history. But by emphasizing what sets us apart from other creatures, we tend to lose sight of the similarities we share with other creatures. In particular, we tend to think (consciously or subconsciously) that it must have taken a lot of time for us to evolve into the special creatures that we know we are, so, therefore, we should see a long, separate evolutionary history for the human lineage.

Nowhere is this issue more apparent then when it comes to the question as to who we are most closely related to among all living things. For a long period of time, this question would be fraught with danger, because the only acceptable answer (at least, in western Judeo-Christian thinking) was that we have no relatives, living or otherwise. According to this dogmatic view, all living things were created by the Creator, so any similarities among living things simply reflect the Creator reusing the same themes—much as an architect might make use of similar elements when designing different buildings. And, we alone among all living things were created in the image of the Creator (so once again we set ourselves apart as being special among all living things—it's hard to get more special than that!); to suggest that we might in fact share common ancestry with other living things would brand you as a heretic and risk a quick and painful visit from the local version of the Inquisition.

But beginning with the Age of Enlightenment, science and reason gradually spread and became accepted to the point where it was possible to entertain other ideas about our relationships with other living creatures. The culmination was Darwin (helped along by numerous other scientists) and the publication in 1859 of his "abstract" on evolution, entitled "On the Origin of Species." This book captured the public imagination like no other scientific work either before or since, selling out completely on the first day of publication. Darwin's ideas were a hot topic in Victorian society, and he was regularly lampooned by cartoonists—a sure sign of fame in those days! Darwin (along with many others) noted the obvious similarities between humans and apes, and the conclusion that humans and apes were descended from a common ancestor gradually became the accepted view in anthropology.

The living (nonhuman) apes (Figure 13.1) can be broadly classified into two groups: the Asian apes, which include gibbons, siamangs, and orangutans; and the African apes, which include gorillas, chimpanzees, and bonobos. Another classification that is sometimes used is based on body size and distinguishes the "lesser apes" (gibbons and siamangs, because of their smaller size) from the "great apes" (orangutans, gorillas, chimpanzees). The features that apes share that

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

FIGURE 13.1

The (nonhuman) apes. Clockwise from top left: gibbons, orangutan, bonobo, chimpanzee, gorilla. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Hylobates_lar_pair_of_white_and_black_02.jpg; https://commons.wikimedia.org/wiki/File:Orang-utan_bukit_lawang_2006.jpg; https://commons.wikimedia.org/wiki/File:Pan_paniscus05.jpg; https://commons.wikimedia.org/wiki/File:Schimpanse_Zoo_Leipzig.jpg; https://commons.wikimedia.org/wiki/File:Gorilla_gorilla_gorilla13.jpg).

distinguishes them from monkeys and other primates include a generally larger body size (except for gibbons and siamangs), absence of a tail, differences in the shoulder joints related to locomotion, larger brain (relative to body size), overall more complex behavior and cognitive abilities, and a longer period of infant development and dependency. And the view that dominated anthropology for a long time as to how humans, African apes, and Asian apes are related—indeed, it was the view I was taught when I was a student—is that depicted in Figure 13.2, which shows that the human lineage is the first to diverge, followed by the divergence of Asian apes from African apes. The date of the human–ape split was put at somewhere between 15 million and as much as 30 million years ago, based largely on fossil evidence dated to about 13 million years ago and ascribed to a creature called Ramapithecus that was reconstructed to be a bipedal ancestor of ours (Pilbeam 1966). So, if you have fossils on the human lineage that are 13 million years old, clearly the divergence of our lineage from the ape lineage has to be older than that. And note that while humans no

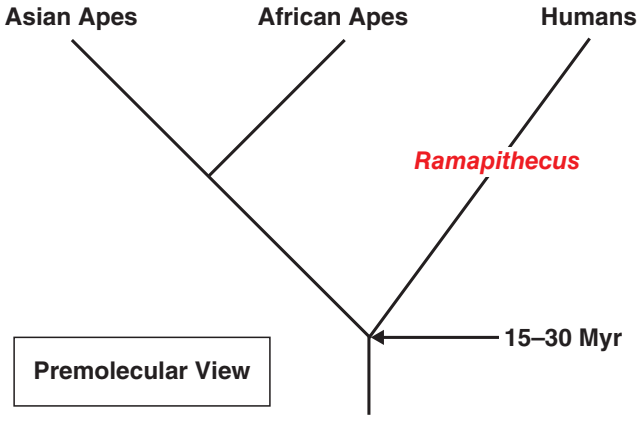


FIGURE 13.2

Premolecular view of the evolutionary relationships of Asian apes, African apes, and humans, showing the presumed 15–30 million year divergence time for the human lineage and the suggested position of Ramapithecus as a human ancestor.

longer occupy the lofty pedestal of having been created in the image of the Creator, still, according to this view we are not so closely related to any other creature. Instead, we've had lots and lots of time as our own separate lineage, enough to evolve into the special creatures that we know we are.

The view of human origins depicted in the tree in Figure 13.2 is based entirely on fossil evidence and studies of comparative anatomy; the first molecular evidence concerning this issue was a study carried out by Morris Goodman (Goodman 1963), one of the pioneers of molecular evolution. Goodman used antigen-antibody reactions to probe the relatedness of species. The basic idea is that the more closely related two species are, the more strongly an antiserum produced against an antigen from one species will react with the other. The results showed that—contrary to the view depicted in Figure 13.2—Asian apes and African apes are not the closest relatives, but rather humans and African apes are the closest relatives. The analysis was qualitative, in that it did not reveal precisely how close is the relationship between humans and African apes, but this first molecular evidence already indicated problems with the then conventional view of the evolutionary relationships of humans and other apes.

Quantitative evidence for the close relationship between humans and African apes was provided by Vincent Sarich and Allan Wilson (Sarich and Wilson 1967a). Sarich, a graduate student working with noted molecular evolutionist Wilson, applied a different type of immunological analysis, called **microcomplement fixation**, that quantifies the differences in particular proteins between different species. The result is a matrix of "immunological distances" (which reflect how different the proteins are) between each pair of species in the analysis, which in this case included humans, African apes, Asian apes, and—as an outgroup—Old World monkeys. Because proteins are encoded by genes, the immunological distance matrix is a kind of genetic distance matrix that can be used to construct a tree (as we saw in Chapter 10), and by assuming that Old World monkeys diverged from apes 30 million years ago (based on fossil evidence), the distances can be converted to time via the molecular clock approach, as described previously. And when Sarich and Wilson did this, they got the astonishing results shown in Figure 13.3, which differed from prevailing wisdom in two important aspects. First, Asian apes do not form a clade, but the great apes do (i.e., orangutans are more closely related to African apes and humans than to gibbons or siamangs). Second, not only are humans and African apes most closely related (in agreement with Goodman's results) but the divergence between humans and African apes was a mere 5 million years ago.

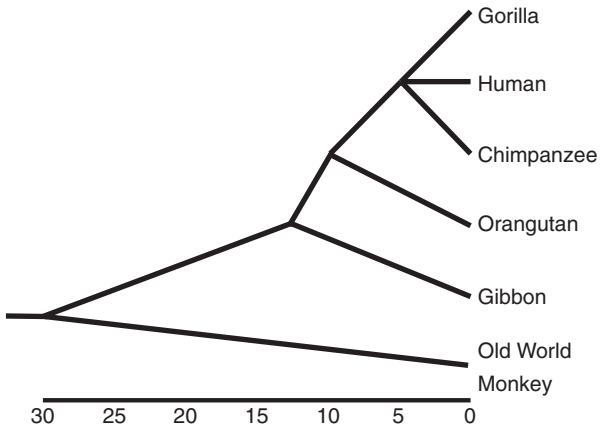


FIGURE 13.3

Diagram illustrating the inferred evolutionary relationships of humans and apes, based on immunological distances (Sarich and Wilson 1967a), and calibrated assuming that apes and Old World Monkeys diverged 30 million years ago. The timescale at the bottom is in millions of years.

Now, scientists are very fond of promoting the view that science is completely objective and dispassionate. In particular, if new results cannot be reconciled with an existing idea about the world, then that idea must be dismissed—no matter how cherished or long-standing—and a new explanation devised to account for the new results. The reality, as any scientist knows, is that it is difficult to overcome ideas that have dominated a field for a long time; instead, there is a tendency to reject the data—and the scientists—which do not fit the theory. And the reception of the results of Sarich and Wilson provides a textbook example, as their results indicating a very close evolutionary relationship of humans and African apes were dismissed out of hand as being too ludicrous to merit serious consideration.

Take the following, for example, courtesy of the anthropologist John Buettner-Janusch: "Unfortunately, there is a growing tendency, which I would like to suppress if possible, to view the molecular approach to primate evolutionary studies as a kind of instant phylogeny. No hard work, no tough intellectual arguments. No fuss, no muss, no dishpan hands. Just throw some proteins into the laboratory apparatus, shake them up, and bingo! We have the answer to questions that have puzzled us for at least three generations" (Buettner-Janusch 1969). It is rather amazing to find a scientist stating so baldly that they would like to suppress other views, not to mention that it is hard to imagine how one could work in a laboratory and not get dishpan hands! Incidentally, Buettner-Janusch later embraced molecular genetic approaches, in particular protein electrophoresis studies of

nonhuman primates. Unfortunately, enterprising students in his laboratory began synthesizing and selling LSD and other illegal drugs; although Buettner-Janusch claimed that this was done without his knowledge, he was sentenced to prison. After his release, embittered and seeking revenge, he sent poisoned chocolates to the judge and the members of the jury that convicted him, as well as to some colleagues who he felt did not support his claims of innocence. Although a few people did get sick, fortunately nobody died; Buettner-Janusch was sent back to prison, where he died a few years later from pneumonia. I have to confess that whenever I read statements exalting how scientists represent the cream of humanity because they are solely motivated by the quest for knowledge, and so forth, my first thought is of scientists like Buettner-Janusch!

The prevailing response to the results of Sarich and Wilson, as expressed in writing, at scientific conferences, and in discussions among colleagues, was that it was all nonsense. Sarich, however, was not one to back down from a fight, and he responded with statements such as: "One no longer has the option of considering a fossil specimen older than ~8 million years a hominid no matter what it looks like" (Sarich 1971). That is, any fossil that predates the molecular date for the divergence between humans and African apes cannot be on the human lineage, no matter how similar it might be in appearance to modern humans—which is, of course, a direct attack on how paleoanthropologists infer where particular fossils fit in the evolutionary scheme of things. My own favorite quote from Sarich about this controversy is the following: "… the biochemist knows his molecules have ancestors, while the paleontologist can only hope that his fossils left descendants" (Sarich 1973). That is, a persistent concern with any fossil is that it may not be a direct ancestor of ours but rather an evolutionary side-branch that went extinct; although, to be fair, while we can be confident that our genes do have an evolutionary past, how one goes about inferring that past from present-day patterns of genetic variation is not so straightforward—indeed, that is a major topic of this book.

To be sure, there was some criticism of the Sarich and Wilson study that focused on the science. In particular, the conclusion of a 5 million year divergence time between humans and African apes rested on the assumption of a molecular clock, and critics rightfully pointed out that if the rate of molecular evolution had not been constant over time, then the divergence time could be considerably older. In response, Sarich and Wilson devised the relative rate test described in the previous chapter and showed that their data did in fact conform to a molecular clock (Sarich and Wilson 1967b). And over the ensuing years, more and more molecular data, first from proteins and then from DNA, showed more and more convincingly that Sarich and Wilson got it right: humans are most closely related to African apes, with a much more recent divergence time than previously suspected (current estimates of the divergence time will be discussed in the "Resolving the trichotomy" section). Which then brings up the question: what about Ramapithecus, supposedly a bipedal ancestor of ours dated to around 13 million years ago? How do we reconcile Ramapithecus with a divergence between humans and African apes of only 5 million years ago or so? It turns out that the supposition that Ramapithecus was a bipedal ancestor of ours was based on just a few teeth and some fragments of the jaw; with further fossil discoveries in the early 1980s, it became apparent that the remains ascribed to Ramapithecus were actually female members of an already described species, Sivapithecus. And since Sivapithecus was thought to be an ancestor of orangutans, the Ramapithecus problem was neatly solved by doing away with it entirely; you will no longer find any mention of Ramapithecus in textbooks, as all fossil remains previously attributed to Ramapithecus are now subsumed under Sivapithecus (there is an old saying that doctors bury their mistakes and paleontologists rename theirs). Incidentally, in case you are wondering how it could be possible to classify fossils of females and males of the same species into different species, take a look at Figure 13.4, which shows the skulls of a male and a female gorilla. These differ dramatically, largely because of the massive crest,

FIGURE 13.4

Skulls of male (left) and female (right) gorillas, showing extreme sexual dimorphism. In particular, note the large crest and much larger canine teeth of the male skull. Reprinted with permission from Wikimedia Commons (http://commons.wikimedia.org/wiki/File:Gorilla_gorilla_skull.jpg).

larger teeth, and overall more robust build of the male skull. Now imagine that you have just a few bits of their skulls and some teeth, and it is easy to see how you might come to the conclusion that fossil remains of male and female gorillas are from different species.

RESOLVING THE TRICHOTOMY

The results of Sarich and Wilson (Figure 13.3) showed that while African apes are our nearest living relatives, the evolutionary relationships of humans, chimpanzees, and gorillas could not be resolved; they formed a **trichotomy**. While it is theoretically possible for three species to diverge simultaneously, in practice it hardly ever happens, so the next big challenge was to determine which two of gorillas, chimpanzees, and humans are most closely related. At first glance, this might seem like the sort of question only a scientist would pose—as the average 2-year-old will tell you, of course chimpanzees and gorillas are most closely related, just look at them (cf. Figure 13.1). But resolving the trichotomy drove a lot of research during the 1980s and 1990s, with people racing to obtain more data and develop new methods. And the astonishing result: it turns out that gorillas, not humans, diverged first, and humans and chimpanzees are most closely related. Indeed, some have called humans the third chimpanzee (after common chimpanzees and bonobos).

The current view of the evolutionary relationships and speciation times for apes and humans, based on near-complete genome sequence data, is shown in Figure 13.5. The speciation times are based on the "conventional" estimate of the mutation rate, which currently is a matter of great debate, as recent studies have suggested a much slower mutation rate—more about this later. As Figure 13.5 shows, the estimated divergence time between humans and chimpanzees is around 4.5–6 million years ago, pretty close to the estimate that Sarich and Wilson obtained more than 40 years ago.

This divergence time estimate does have implications for interpreting the fossil record—as Vince Sarich pointed out many years ago, anything older than this divergence cannot, by definition, be on the human lineage (no matter what it looks like). In particular, the oldest fossils that have been attributed to the human lineage are *Sahelanthropus tchadensis* from Chad (dated to 6–7 million years ago) and *Orrorin tugenensis* from Kenya (dated to about 6 million years ago). While these dates do (just) fit within the upper limits of the human–chimpanzee divergence, it is by no means universally accepted that these fossils are of human ancestors; others have argued that one or both may be ancestors of humans and chimpanzees, ancestors of gorillas, or an extinct ape that was not ancestral to any of the living African apes. Unfortunately, the public (and funding agencies) are quite keen on discovering the fossils of the earliest *human* ancestors and much

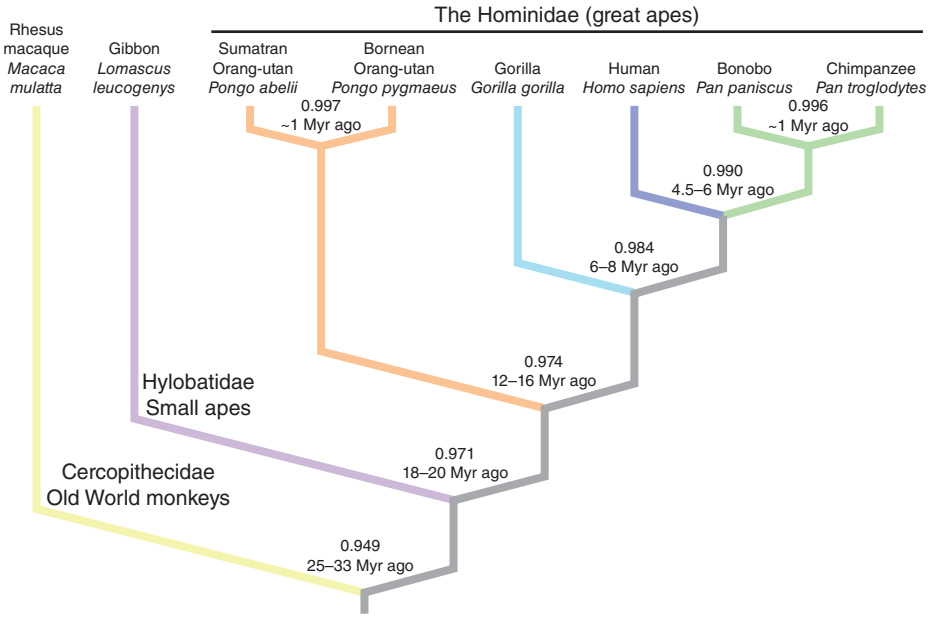


FIGURE 13.5

Evolutionary tree based on genome sequences, showing sequence identity and estimated divergence times (using the phylogenetic rate estimate). Reprinted with permission from Locke, D., et al., "Comparative and demographic analysis of orangutan genomes," *Nature* 469:529, 2011.

less keen on fossils of other ape lineages, so there is a natural tendency for fossil-hunters to interpret their finds as our ancestors (and, of course, if you can claim that your new fossil changes everything we thought we knew about human evolution, so much the better in terms of publicity!).

How important has the molecular evidence been for understanding how we are related to other apes? While there is no way of knowing for sure what our view would be on the evolutionary relationships of humans and apes if we had only fossil evidence, and no molecular evidence, arguably it is highly unlikely that we would have arrived at the relationships shown in Figure 13.5. Instead, it is quite likely that humans would still be viewed as only distantly related to other apes—indeed, in the absence of molecular evidence, I think it quite probable that something like Figure 13.2 would still be the predominant view. In my opinion, the demonstration of an extraordinarily close evolutionary relationship between humans and chimpanzees is perhaps the most significant contribution of the molecular approach to anthropology, as I doubt very much that we would have arrived at that conclusion in the absence of the molecular evidence.

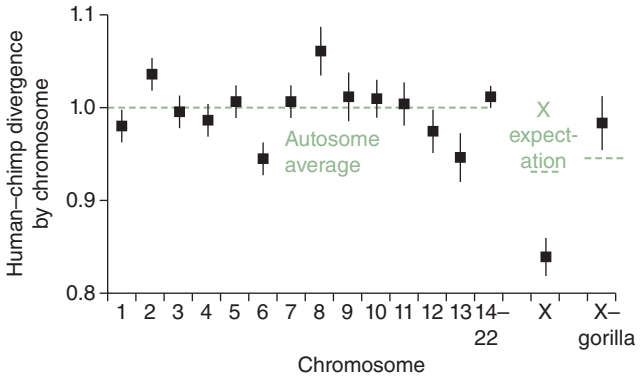


FIGURE 13.6

Sequence divergence between humans and chimpanzees per chromosome, showing that the human–chimpanzee divergence for the X chromosome is much smaller than the autosomal average or the expectation for the X chromosome (dashed lines). By contrast, the human–gorilla sequence divergence on the X chromosome is similar to the expected value. Reprinted with permission from Patterson, N., et al., "Genetic evidence for complex speciation of humans and chimpanzees," *Nature* 441:1103, 2006.

COMPLICATIONS

While the evolutionary scenario depicted in Figure 13.5 may seem neat and tidy, there are a number of complications hidden under the surface. First, while overall humans are more closely related to chimpanzees (98.63% DNA sequence similarity) than gorillas (98.25% DNA sequence similarity), this is not true for the entire genome. As already mentioned in the previous chapter, for about 30% of the genome, either humans or chimpanzees are more closely related to gorillas than to each other. This may seem astonishing, but in fact it is a natural consequence of ancestral polymorphism resulting in incomplete lineage sorting, as discussed previously. Given the short time period between the divergence of the gorilla lineage from the human–chimpanzee ancestor and the divergence of the human and chimpanzee lineages (Figure 13.5), it is to be expected that a large number of polymorphisms in the ancestral gorilla–chimpanzee–human population will also still be polymorphic in the ancestral chimpanzee–human population, thereby providing an opportunity for shared gorilla–chimpanzee or shared gorilla–human genomic segments.

Moreover, while the average divergence between a human and a chimpanzee chromosome is about 1.4%, the X chromosome shows a radically different pattern; human and chimpanzee X chromosomes are much more similar than are other chromosomes (Figure 13.6). To be sure, somewhat greater similarity for the X chromosome is to be expected. This is because in a population with equal numbers of males and females, there are three X chromosomes for every four autosomes in the gene pool, so the effective size for the X chromosome is 3/4 that of the autosomes. This reduction in effective size is expected to lead to a corresponding reduction in the time to the most recent common ancestor, as discussed in the previous chapter. Moreover, because the average X chromosome spends twice as much time in females as in males, and the mutation rate is lower in females than in males (probably because there are more cell divisions, and hence more opportunities for errors during DNA replication, in the male germ line than in the female germ line), fewer mutations are expected on the X chromosome than on the autosomes.

But there appears to be more similarity between human and chimpanzee X chromosomes, relative to autosomes, than can be explained by these factors. In addition, there are many fewer X chromosomal than autosomal segments that show incomplete lineage sorting (i.e., where either the human or the chimpanzee sequence is more closely related to the gorilla sequence). These observations of unusually low divergence between the human and chimpanzee X chromosomes led population genomicists Nick Patterson and David Reich to propose a complex speciation model for humans and chimpanzees (Patterson et al. 2006), in which humans and chimpanzees first started diverging before 6 million years ago, then after some period

of isolation (perhaps a million years or so) there was further genetic exchange (i.e., hybridization) between human and chimpanzee ancestors, followed by permanent separation and complete speciation. And why should this lead to unusually low divergence specifically on the X chromosome? Patterson and Reich proposed that incompatibilities in hybrids between the X chromosome of one parental species and the autosomes could quickly lead to the elimination of that X chromosome via selection; in support of this idea, it is known that genetic barriers to hybridization between species often involve the X chromosome. So, according to this hypothesis, during the genetic exchange between the somewhat diverged human and chimpanzee ancestors, the "original" X chromosome in the human lineage (which carried the expected amount of divergence) was replaced by an X chromosome that was more closely related to the chimpanzee X chromosome. However, this complex speciation model is quite controversial, and other explanations have been proposed for the lower divergence of the X chromosome—for example, male versus female mutation rates may have been different in the past. Or, selection on genes on the X chromosome may also play a role—it is known that selection involving recessive alleles is more efficient on the X chromosome than on autosomes, because the hemizygosity of the X chromosome in males means that males will exhibit the phenotype of any X-linked recessive allele, and hence selection can act on the trait. By contrast, as we saw back in Chapter 5, selection on an autosomal recessive allele is less efficient, because selection requires homozygotes for such alleles, but homozygotes will be very rare in the population until the allele reaches appreciable frequencies. Whether or not these other factors can fully explain the data is still a matter of debate; the current state of affairs is that the complex speciation hypothesis is just so weird that it seems like there must be some other explanation, but what that might be is still a mystery. Nevertheless, while the explanation for the lower X chromosome divergence between humans and chimpanzees remains controversial, it is quite clear that there is more to the human–chimpanzee speciation event than meets the eye.

Finally, the divergence times depicted in Figure 13.5 are based on an estimate of the mutation rate of about 1×10^{-9} substitutions per base-pair per year. This estimate, in turn, comes from various calibrations that rely on fossil evidence. For example, the divergence time between Old World monkeys and apes has been "confidently" dated to 25–30 million years ago, based on fossils ascribed to both lineages that are at most 23 million years old (note, however, that this 5 million year time range in the divergence time estimate is about the same as the amount of time that the human lineage has been in existence!). The rhesus macaque genome sequence differs from the human genome sequence by about 6.46%, so if you do the math (0.0646 sequence divergence = 0.0323 sequence evolution per lineage, divided by 25–30 million years), you end up with 1.1–$1.3 - 10^{-9}$ substitutions per base-pair per year.

As discussed in the previous chapter, such estimates of the molecular clock rate, based on comparison of the amount of sequence divergence between different species, are known as phylogenetic rate estimates. However, there are several recent studies that conclude that the mutation rate is actually lower by about half (reviewed in Scally and Durbin 2012). These studies are based on next-generation sequencing of parents and children in order to identify new mutations in the children that are not present in the parents and hence provide a direct estimate of the mutation rate per generation. This difference between the pedigree rate estimates (coming from family data) and the phylogenetic rate estimate is even more puzzling when you consider that ordinarily you would expect the pedigree rate to be higher, not lower, than the phylogenetic rate. This is because not all of the new mutations observed in children will survive in the population to be counted as fixed differences between species; some proportion of these new mutations will go extinct either because of drift and/or because they are deleterious and hence selected against. In fact, for mtDNA, the pedigree rate estimates are higher than the phylogenetic rate estimates, as expected, so why this isn't also the case for autosomal DNA data is, at the moment, mysterious.

The consequences for this discrepancy for molecular clock estimates of divergence times are clear enough: since the divergence time and mutation rate are linearly related, lowering the mutation rate by half would double the divergence times in Figure 13.5, which raises all sorts of complications. On the one hand, the human–chimpanzee divergence would be pushed back from 4.5–6 to about 9–12 million years, which would easily accommodate those who argue that *Sahelanthropus tchadensis* and/or *Orrorin tugenensis* is on the human lineage. However, the divergence of orangutans from other apes then becomes at least 24 million years, which contradicts fossil evidence younger than this date from what are thought to be ancestors of all of the great apes (such as *Proconsul*, around 19 million years ago). Of course, one of the take-home lessons from the *Ramapithecus* debacle is that one should be extremely cautious about rejecting molecular clock dates on the basis of fossil evidence! Still, applying the lower, family rate estimate to other parts of the primate (and mammalian) phylogeny raises more problems than it solves.

At the time that I write this, there is no consensus as to how to resolve this discrepancy or as to what rate should be used. One potential complication with the pedigree rates is that an estimate of the generation time is needed to convert what is observed, namely,

mutations per generation, to rates in mutations per year. As mentioned previously, typical estimates for generation times in humans are 25–30 years, but if the generation time was much shorter in the past, say 10–15 years for a few million years after humans and chimpanzees diverged, then the pedigree and phylogenetic rate estimates would be similar. However, a recent study of chimpanzees and gorillas, based on census data from wild populations, came up with similar generation time estimates (25 and 20 years, respectively) as those for humans (Langergraber et al. 2012). So if the generation time was indeed much lower around the time of human–chimpanzee speciation, then the generation time apparently would have increased independently in the lineages leading to humans, chimpanzees, and gorillas, which hardly seems likely—it is much more parsimonious to suppose that the generation time has always been on the order of 20–25 years or so, at least since the human–chimpanzee–gorilla divergence.

Another possible explanation that has been suggested is a recent slowdown in the rate of molecular evolution. There is ample evidence to suggest that the rate of molecular evolution has decreased in general in primates and also in apes relative to Old World monkeys (Li and Tanimura 1987; Steiper et al. 2004). Why this is the case is not known but may be related to metabolic rates. In general, the faster the metabolic rate, the higher the rate of molecular evolution—possibly because higher metabolic rates may lead to higher mutation rates (which gives new meaning to the old saying about living fast and dying young!). But this is far from settled; rates of molecular evolution are also correlated with generation time and body size, among other things, and both of these are also correlated with metabolic rate. So, what is cause and what is effect is still hotly debated. Nonetheless, it is certainly the case that apes are larger, have longer generation times and slower metabolic rates than do Old World monkeys, and any or all of these may explain (or contribute to) the slower rate of molecular evolution in apes. Regardless of the underlying cause, it has been proposed that faster rates of molecular evolution in the past, followed by a more recent slowdown could potentially help reconcile the disparity between phylogenetic and pedigree estimates of the mutation rate, although whether this is really the case needs further investigation.

APE GENETICS AND GENOMICS

The fact that humans are apes means that there is much we can potentially learn about our evolutionary past from genetic and genomic studies of other apes. However, in doing so, it is important not to fall into the trap of thinking that chimpanzees (or other apes) are our ancestors; chimpanzees have had just as much time to evolve from the common human–chimpanzee ancestor as we have had. Therefore, the traits and behaviors that chimpanzees exhibit may not be at all representative of those of the human–chimpanzee ancestor. Still, comparisons of our genome to that of chimpanzees (and other apes) can help identify the genetic changes that were important in our own evolution, and this will be covered in Chapter 17. Moreover, using the same tools and methods as described previously for studying human genetic diversity, studies of genetic diversity in other apes can shed light on their evolutionary and demographic history. The various species of apes also exhibit tremendous diversity in many traits, in particular social structure: gibbons form (mostly) monogamous pairs; orangutans are largely solitary; gorillas live in family groups consisting of (usually just) one dominant male and a harem of several females, along with their offspring and subadult males; chimpanzees live in multimale, multifemale groups centered around male dominance; and bonobos live in multimale, multifemale groups centered around female dominance. Genetic studies can help elucidate the causes and consequences of such diversity, and genetic studies of wild primate populations have been greatly enabled by the polymerase chain reaction and other methods that permit noninvasive sample collection (feces turn out to be an excellent source of DNA, even though the people who collect and process fecal samples have to put up with the inevitable bad puns about the shitty nature of the work). As just one example, genetic studies of paternity have proven useful in supplementing field observations concerning who is fathering the offspring in a group. It turns out, for example, that the dominant, silverback male in a gorilla group does not always father all of the offspring (Bradley et al. 2005)—further proof that humans are not so different from our ape relatives when it comes to traits such as nonpaternity.

A thorough description of primate genetics and genomics is beyond the scope of this book—indeed, it deserves its own textbook. However, in the chapters to come, we will occasionally make reference to primate genetic studies that are especially relevant to particular topics concerning humans, such as the impact of residence patterns on genetic variation. Here, it is worth pointing out one aspect of genetic diversity in which humans do differ from other apes: humans have the lowest genetic diversity of any ape. This was first noticed with mtDNA studies (Figure 13.7) and has since been confirmed with genome sequence data (Table 13.1—although, to be sure, bonobos run us a close second in genomic diversity). Table 13.1 also gives estimated census population sizes for humans

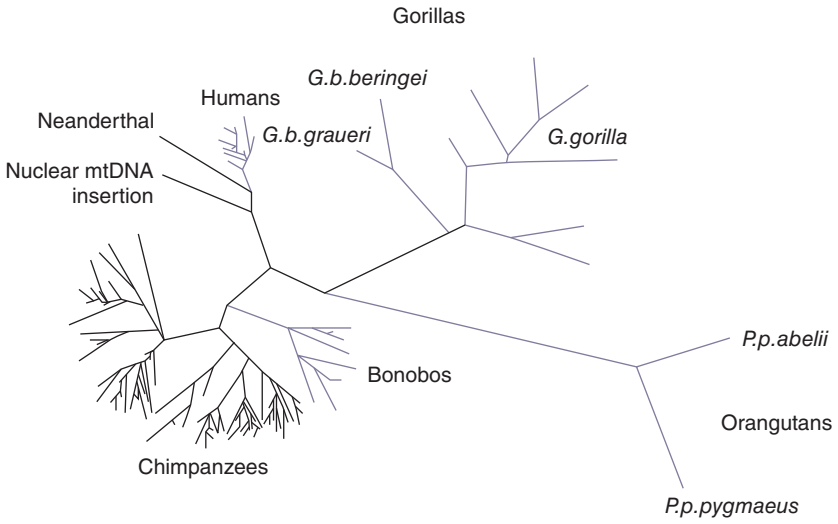


FIGURE 13.7

Tree of mtDNA hypervariable segment 1 sequences from great apes, including sequences from a Neanderthal and from an insertion of mtDNA found in the nucleus. The lengths of the branches are proportional to the number of mutations on that branch. Note the extremely short branches in humans that cluster tightly, compared to much longer and more dispersed branches for the great apes, indicating much greater mtDNA diversity in great apes than in humans. Reprinted with permission from Gagneux, P., et al., "Mitochondrial sequences show diverse evolutionary histories of African hominoids," *Proceedings of the National Academy of Sciences USA* 96:5077, 1999.

TABLE 13.1 ■ Genetic diversity (number of polymorphic sites per 1000 base-pairs, based on complete genome sequences) and current estimates of census population size for great apes and humans. Genetic diversity estimates are from Prado-Martinez et al. 2013; census size estimates for the great apes are from the World Wildlife Fund (wwwf.panda.org)

Species	Diversity	Census size
Orangutan	0.46	60,000
Gorilla	0.84	100,000
Chimpanzee	1.17	150,00–250,000
Bonobo	0.38	<15,000
Human	0.37	7 billion

and other great apes; note that population sizes are much larger for humans than for any other ape. In fact, there are fewer wild great apes in the world than there are people in the city I live in (Leipzig, Germany, which is hardly a large city). And yet, all of these apes have more genetic diversity than us humans. This point was already made back in Chapter 3, in which we discussed effective population size estimates for humans and apes but is worth making again here: compared to our nearest living relatives, humans are characterized by unusually low levels of genetic diversity, which probably indicates that we had very small population sizes (even smaller than current ape populations) at some point in our evolutionary past.

CHAPTER 14

THE ORIGINS OF OUR SPECIES

The previous chapter covered the genetic evidence for a close evolutionary relationship between humans and chimpanzees, which is one of the most important contributions of the molecular approach to anthropology. Now we will turn to another major question where molecular anthropology has provided key insights: namely, how did our own species originate? This question is actually more complicated than it appears at first glance, because it looks as if we are interested in the origin of a single entity (us). In fact, when we look at humans around the world, there seems to be an incredible diversity in their physical appearance (Figure 14.1), and so when we ask how did our species originate, what we would really like to know is, how did all of this variation arise, and is it ancient or recent?

Looking at pictures of people from around the world, such as shown in Figure 14.1, inevitably raises the question as to what this variation actually means. In particular, can we think of these people as representing different races? Whether or not there is any basis for applying the concept of "race" to humans is one of the most controversial topics in anthropology—indeed, in human society. To put it simply, are there different human races? And if so, what are they, and how do they differ? Before turning to our origins, let's first see what genetics has to say about the existence of races.

If we're going to talk about race, we should start by defining what we mean, since a lot of confusion can be avoided by making sure that everyone is talking about the same thing. Here are some definitions of "race," compiled by Jeff Long and Rick Kittles (2003):

> A great division of mankind, characterized as a group by the sharing of a certain combination of features, which have been derived from their common descent, and constitute a vague physical background, usually more or less obscured by individual variations, and realized best in a composite picture.
>
> … an aggregate of phenotypically similar populations of a species, inhabiting a geographic subdivision of the range of a species, and differing taxonomically from other populations of the species.
>
> Races are genetically distinct Mendelian populations. They are neither individuals nor particular genotypes, they consist of individuals who differ genetically among themselves.
>
> A subspecies (race) is a distinct evolutionary lineage within a species. This definition requires that a subspecies be genetically differentiated due to barriers to genetic exchange that have persisted for long periods of time; that is, the subspecies must have historical continuity in addition to current genetic differentiation.

You begin to see the problem—none of these definitions are precise enough to be useful in attempting any sort of racial classification of humans. What is meant by "great divisions of mankind," "aggregates of phenotypically similar populations," "genetically distinct Mendelian populations," "distinct evolutionary lineages," and so forth? All of these are subjective and open to interpretation (and misinterpretation).

So, keeping in mind that it is difficult to come up with a precise definition of "race," let's take a quick look at the history of how the concept of race has been applied to humans. One of the most influential scientists to weigh in on this topic was the great Swedish naturalist Carl Linnaeus. Linnaeus was responsible for developing the familiar "binomial" nomenclature for designating species, consisting of the genus name (e.g., "*Homo*"), followed by the species name (e.g., "*sapiens*"). Linnaeus is thus considered to be the father

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

FIGURE 14.1

Human diversity from around the world. Reprinted with permission from Wikimedia Commons (clockwise from top left: https://commons.wikimedia.org/wiki/File:Ales_hrdlicka.jpg; https://commons.wikimedia.org/wiki/File:Trugannini_1866.jpg; https://commons.wikimedia.org/wiki/File:Huichol_indian.jpg; https://commons.wikimedia.org/wiki/File:Yukawa.jpg; https://commons.wikimedia.org/wiki/File:Sardar_patel_%28cropped%29.jpg; https://commons.wikimedia.org/wiki/File:Bedouinwomanb.jpg; https://commons.wikimedia.org/wiki/File:Asaro_Mud_Man_Kabiufa_PNG.jpg; https://commons.wikimedia.org/wiki/File:San_tribesman.jpg).

of taxonomy and became one of the most well-known scientists of his time (and beyond)—in recognition of his accomplishments, in 1959, he himself was designated the type specimen (i.e., the specimen used to first describe the species) for *Homo sapiens*, a singular honor indeed!

In the 1758 edition of his classic work, *Systema Naturae*, Linnaeus described and classified more than 12,000 species of animals and plants. Most of his classifications were based on physical descriptions, geographic distributions, and so forth, but rather peculiarly, when it came to humans, he included moral judgments in his classification. Linnaeus identified four different races of humans: in the New World there is *Homo sapiens americanus* ("red, ill-tempered, subjugated … Obstinate, contented, free … Ruled by custom."); Asia is the home of *Homo sapiens asiaticus* ("yellow, melancholy, greedy … Severe, haughty, desirous … Ruled by opinion."); in Africa, we find the lowly *Homo sapiens afer* ("black, impassive, lazy … Crafty, slow, foolish … Ruled by caprice."); and in Europe, we find the pinnacle (of course), *Homo sapiens europaeus* ("white, serious, strong … Active, very smart, inventive … Ruled by laws.").

It would be easy to dismiss Linnaeus' characterization of the various human races as simply reflecting the Eurocentric world view of that time—easy but not entirely correct. Because at around the same time another naturalist, Georges-Louis Leclerc (also known as the Comte, or Count, de Buffon, after his parents came into a sizable estate and his father became Lord Buffon) was writing in a very different way about human races. In his 1749 essay, *Varieties of the Human Species*, Leclerc was not interested in merely classifying humans into different categories. Instead, he set out to describe the range of physical and cultural variation that exists across the human species. Linnaeus was content to simply ask how many categories (races) of humans are there and what are they; Leclerc wanted to know how variation is patterned in humans and how it got to be that way.

Leclerc was, in many ways, ahead of his time. For example, in writing about African slaves (which at that time were the only Africans that most Europeans observed directly), in his magnum opus *Histoire Naturelle* (an encyclopedic work of 36 volumes published from 1749 to 1789), he had this to say:

> I cannot write their history without describing their state … They are forced to labor, and yet commonly are not even adequately nourished. It is said that they tolerate hunger easily, that however little they eat or sleep, they are always equally tough, equally strong, and equally fit for labor. How can men in whom there rests any feeling of humanity adopt such views?

This is not to say that he was completely enlightened—overall, he was convinced that Europeans were superior to Africans—but, as the aforementioned statement indicates, he at least realized that the appalling conditions under which African slaves lived and labored might have some impact on their existence and how they were perceived by others. Anyway, the overall achievement of Linnaeus' monumental work in arriving at a classification of everything—and his subsequent elevation in stature to the premier naturalist of his time—meant that his views eclipsed those of others, and Leclerc was relegated to relative obscurity. The Linnaean view that there are indeed racial differences among different groups of humans (which, after all, corresponded to prevailing notions of the time) became entrenched in biology and, subsequently, anthropology.

What does genetics have to say about the existence of human races? Confusingly, genetic data have been argued to both refute and support the concept of human races. On the one hand, there is the well-known result (cf. Table 10.2) that only about 15% of the genetic variance in the total human species can be attributed to differences among populations, and hence 85% reflects differences among individuals from the same population. This would certainly seem to support the view that overall, the genetic similarities among human populations far outweigh the differences—and hence, argue against any genetic support for the concept of human races. On the other hand, as we saw back in Chapter 11, we can apply clustering methods to the same genetic data and easily distinguish among human populations from different continents (cf. Figure 11.22). If you equate populations from different continents to different races, then these results would certainly seem to support the view that the genetic epidemiologist Neil Risch espoused in 2002: "The greatest genetic structure that exists in the human population occurs at the racial level" (Risch et al. 2002).

So how can we reconcile these two views of the genetic data? The answer is that these contrasting approaches to the genetic data are actually asking different questions about the data, which in turn leads to (apparently) different answers. The first approach is akin to the Leclerc view and asks how is genetic variation patterned in the human species? And the answer is that most of the genetic variation is shared by individuals from different populations, which does not support the concept of races. The second approach is more like the Linnaean view and wants to know what are the categories that we can place humans into? And the answer is that if we use an approach that assumes that we can indeed place humans into categories based on genetic data, then the categories that best fit the genetic data are continental-level groupings (which for many people would correspond to races).

For those of you who know something about statistics, it is like the difference between analyses that describe the overall patterns of variation based on all of the data, versus discriminant analyses, which focus on just those aspects of the data that can be used to assign individuals into categories. Let's use a completely made-up example to make this clear. Suppose my categories are the four races of Linnaeus (native Americans, Europeans, Asians, and Africans). Furthermore, suppose there is only one variable nucleotide position in the entire genome of all individuals, and at this one variable position all native Americans have an A, all Europeans have a C, all Asians have a G, and all Africans have a T; all of the remaining 3 billion nucleotide positions are invariant, with the same nucleotide in all individuals. Clearly, with only one variable nucleotide position out of 3 billion, all people are then extraordinarily similar genetically (99.9999997% identical, to be precise!), and any supposed "racial" differences would not reflect this overall genetic similarity. But just as clearly, I could easily assign the "race" of any individual by genotyping just the one variable nucleotide position—if all I am interested in is putting people into categories, then I only need to know about the variation that lets me do so, and I can ignore all the rest. Different questions, different answers.

In reality, we can define a set of categories based on broad geographical groupings, such as continent of origin, and then find genetic markers that will allow us to place people into those categories with a fair degree of accuracy—this is the whole premise behind so-called **ancestry-informative markers** (or AIMs for short), which are used to classify the ancestry of individuals (e.g., European, African, east Asian, etc.), based on predefined categories. Ancestry-informative markers are very useful for placing people into these various ancestry-defined categories of interest—but again, this fact begs the question as to whether or not the existence of these categories (i.e., races) is supported by overall patterns of genetic variation. We can even undertake analyses (such as STRUCTURE) where we don't predefine categories but see instead what categories come out of the data, but again such analyses start with the supposition that categories do indeed exist. They do not tell us how much support these categories receive from all of the data (i.e., is it just one nucleotide out of 3 billion, as in the aforementioned made-up example)?

What sort of analysis can we do to investigate whether racial (or other) categories exist in genetic data, without first presupposing that there are indeed categories in the data? Here it gets more difficult, as

you first have to think of some prediction that the existence of races would make regarding overall patterns of genetic variation, and this in turns leads to lots of arguments as to whether a racial view would actually lead to the particular prediction in question. But we can at least try. One prediction that seems hard to argue with is that if racial categories are evident in overall patterns of human genetic variation, we should find genetic boundaries between these categories. In other words, we should be able to detect "breaks" in the distribution of genetic variation that correspond to the boundaries between these racial categories. So, let's look at genetic differences between populations versus the geographic distance between them and see whether there are indeed any breaks that would correspond to racial categories. Figure 14.2 shows just such an analysis, based on the HGDP populations (described in Chapter 9), which comprise a worldwide sampling of 52 human populations. There are two versions of the plot of genetic versus geographic distance in this figure; look first at the black and white version. It is quite clear that there are no obvious breaks in the distribution of genetic versus geographic distances—you cannot tell from this figure which are the genetic distances between groups from the same continent and which are the ones from different continents. The color version shows that genetic distances between groups from the same continent are mostly less than the genetic distances between groups from different continents, which at first glance might support the view that different continental groups of humans are races. However, the geographic distances between groups from different continents are also (mostly) larger than the geographic distances between groups from the same continent. What happens if we, therefore, control for geographic distance? One way we can do this is to examine pairs of groups separated by the same geographic distance and see whether those groups from the same continent are more similar genetically than groups from different continents. The result that is evident in Figure 14.2 is that groups that are separated by the same geographic distance show more or less the same genetic distance, regardless of whether they come from the same continent or from different

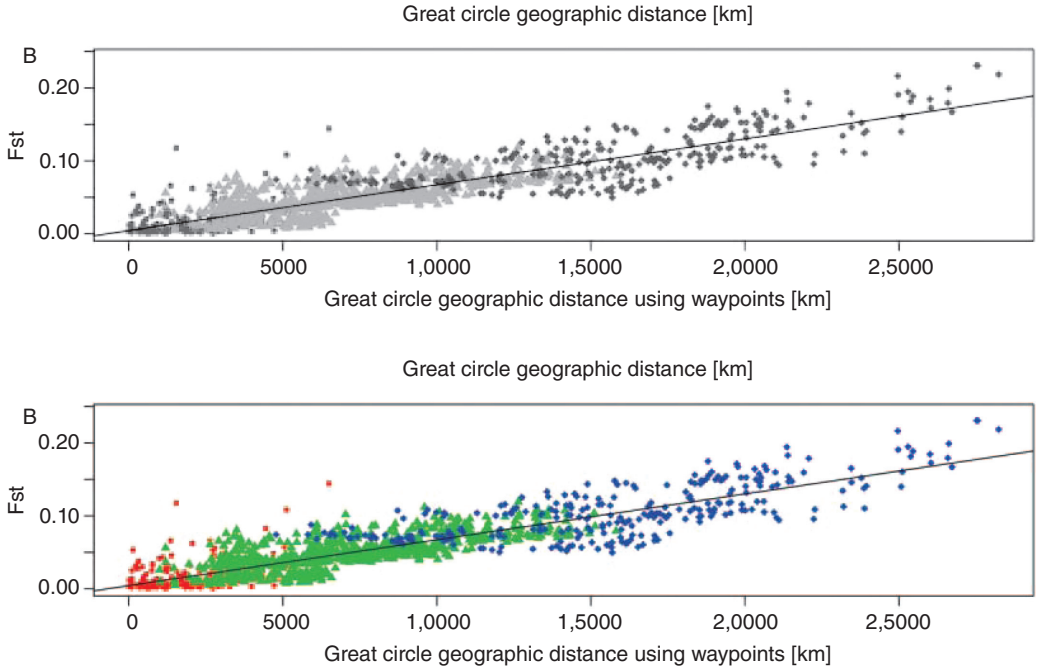


FIGURE 14.2

Genetic (F_{ST}) versus geographic (X-axis) distances for all pairs of populations in the CEPH–HGDP, based on 783 STR loci. The geographic distances allow for the curvature of the earth (great-circle distances) and go around major geographic obstacles to human migration (such as the Caspian Sea or the Himalayas) rather than directly across such obstacles, by making use of waypoints. The top and bottom figures are the same, except that the bottom figure is colored such that red points are for pairs of populations within the same continental region, green points are between African and Eurasian populations, and blue points are between the Americas/Oceania and all others. Modified with permission from Ramachandran, S., et al., "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa," *Proceedings of the National Academy of Sciences USA* 102:15942, 2005.

continents—there is no clear distinction between intercontinental versus intracontinental comparisons when one controls for geographic distance. Hence, genetic distances between groups (largely) reflect their geographic distance, and there is no evidence of significant breaks in the distribution of genetic distances that would correspond to racial categories. Moreover, the way in which populations are sampled can influence the definition and assignment of categories (as pointed out in Chapter 11, especially Figure 11.23).

So, to summarize, we can readily define various categories of humans that would correspond to some views of what races are, and we can find genetic markers that would allow us to assign most people to one of these categories with a fairly high degree of accuracy. We can also find significant differences in allele frequencies in people from different geographic regions of the world. But, overall patterns of genetic variation are primarily and strongly influenced by geography, with no clear breaks in the distribution of genetic versus geographic distances that would correspond to races. There thus seems little reason to pursue the Linnaean approach of asking what labels we can put on various groups of people; instead, in the remainder of this book, we will follow Leclerc and focus on asking, what are the patterns of genetic diversity and how did they get to be that way? Unfortunately, there are those who persist in thinking that the concept of human races remains a useful way to think about human genetic variation—just recently, I was asked to help edit a special issue of a journal on the topic "Genetics of Human Race" (!), which I declined. Even more appalling is the recent book entitled "A Troublesome Inheritance: Genes, Race, and Human History," by former New York Times science writer Nicholas Wade. Wade takes the view that selection has influenced variation in intelligence and social behavior in different populations, and one can find statements in his book to the effect that it is inherent in their nature that Europeans prefer open societies and the rule of law, while Chinese prefer political hierarchy and conformity, and Africans suffer unfortunate tendencies concerning work ethics (or the lack thereof) and propensity to violence. To paraphrase baseball philosopher Yogi Berra, it's like Linnaeus, all over again! In response, several prominent human evolutionary geneticists circulated a letter decrying the misuse of results from population genetics by Wade in his book; this letter was ultimately signed by more than 130 other scientists (including, I'm happy to say, myself) and published in the New York Times.

Still, there are those who would argue that there are indeed situations in which thinking in terms of races can be useful. For example, there are well-documented differences in how rapidly or completely people of European, African, or Hispanic ancestry metabolize certain drugs. Therefore, when setting the initial dose level (or duration of treatment) with such a drug, physicians will commonly take into account the ethnicity (i.e., race) of the patient. And while on the one hand this makes sense, on the other hand, it is a rather poor proxy for the ideal situation, in which one would tailor the drug treatment to the specific patient. For some of the documented "racial" differences in drug metabolism, the underlying mechanism has been investigated and shown to involve a specific genotype at a specific drug-metabolizing gene. Given that there can be variation in different populations in allele frequencies at such genes, either via demographic history or via selection, the result can be differences in drug metabolism in people from different populations. But rather than rely on the population affiliation of the patient, it would actually make more sense to genotype patients for such genes, as the concordance between "race" and genotype is never 100%. In other cases, the underlying genetic mechanism behind "racial" differences in drug metabolism is either more complicated or may involve environmental as well as genetic influences (after all, European–Americans, African–Americans, and Hispanics hardly all have the same environment, even if they all live in the same city and go to the same hospital). In any case, what seems most logical is to factor in "race" in treatment with drugs only in the absence of any further specific knowledge about why drug metabolism varies. The ultimate goal should be tailoring treatments to the actual specific genetic (and environmental) conditions of each patient, rather than prescribing a treatment based simply on a patient's self-described "race."

HUMAN ORIGINS: THE FOSSIL RECORD

Before there were genes there were fossils, and for a long time the fossil record was the only source of information about the evolution of our species. Yet, it is difficult to come up with a definitive physical description of what sets us apart as a species from our predecessors in the fossil record. This is to some extent to be expected, because the concept of different species will always be imprecise as it attempts to impose a categorical distinction on what is an inherently continuous process of differentiation. Moreover, the recognition of different species from fossils rests on assessments of how physically different they are, which can be quite difficult and subjective—recall the example of Ramapithecus from the previous chapter. Still, fossil evidence has been and continues to be an important source of information concerning human evolution, and there are many aspects of our evolutionary past that are revealed only in the fossil record, so even though this is a molecular anthropology textbook, a

brief overview of the fossil evidence for human origins is warranted.

As we saw in the previous chapter, our lineage split from that leading to chimpanzees somewhere around 4.5–10 million years ago (taking into account the current uncertainty around the mutation rate discussed in the previous chapter). The major differences between humans and chimpanzees that would show up in the fossil record include bipedal locomotion (leading to all sorts of associated changes in the postcranial skeleton), smaller teeth, and bigger brains of humans than chimpanzees. And, it is generally assumed that the human–chimpanzee ancestor was much more chimp-like than human-like in these characteristics (even though chimpanzees have had just as much time to evolve from the human–chimpanzee ancestor as we have had). So, any fossils that appear to be more human-like than chimp-like in any of these characteristics would be assigned as potential ancestors of ours. Moreover, since bigger brains would seem to be the most important of these characteristics, the experts generally thought that this is what should change first in the human lineage, and so therefore early fossils on our lineage should show increases in brain size. Indeed, this preconception was so strong among anthropologists that it led many to reject one of the earliest fossil finds of a true ancestor of ours, the Taung child (discovered in South Africa in 1924), because it had small teeth but a chimpanzee-sized brain, but accept as authentic a forgery, the Piltdown skull (Figure 14.3). The Piltdown skull had apelike teeth combined with a big brain (because some enterprising person had combined an orangutan jaw with a modern human skull, skillfully filing down the teeth and staining the skull and jaw to make the find appear authentic). Discovered in 1912, the Piltdown skull was also accepted as authentic because it was found in Britain, where it was, of course, naturally assumed that humans would have evolved, not some "lowly" place like Africa. Because the Piltdown skull fit so well with these preconceptions, it took more than 40 years before it was finally accepted to be a hoax.

From the time of the human–chimpanzee divergence until about 2 million years ago or so, all of our evolution occurred in Africa. There are a number of fossils from this time period generally classified as australopithecines (which means "southern ape"), found in southern and eastern Africa, which tend to be characterized by increasing evidence of bipedal locomotion and smaller teeth—increases in brain size come later. Figure 14.4 shows some of the inferred species and associated ages; how many species there actually are, and which of these are actually our ancestors and which are more likely to be extinct side branches, continues to be a source of great debate. Rather than focusing on this rather narrow question, which probably cannot be answered anyway given the fragmentary nature of the fossil record, it would seem to make more sense to focus on general trends. And the dominant trends during the period of the australopithecines seems to be increasing specialization for bipedal walking and reductions in tooth size (the latter probably reflecting changes in diet), with major increases in brain size only coming later with the appearance of our genus, *Homo*. However, we should also keep in mind that there is a natural tendency for us humans to focus on those aspects of australopithecines that associate them with us; if we instead imagined that we were examining these fossils from the viewpoint of chimpanzees, we might conclude that australopithecines were simply chimpanzees that had a funny way of walking.

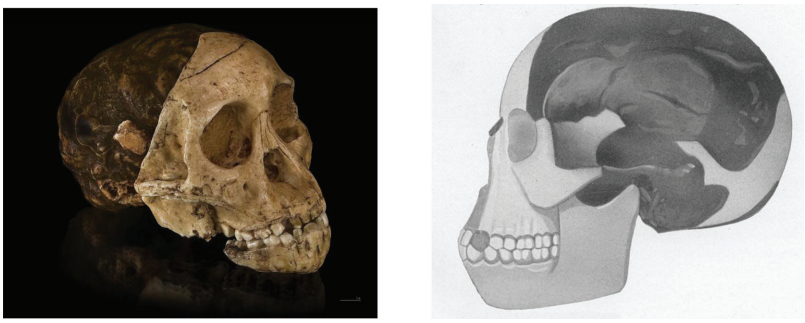


FIGURE 14.3

Fossils of the Taung child (left) and Piltdown man (right). The authentic Taung child fossil was originally thought to not be an ancestor of humans because it had a small brain (counter to expectations), whereas the fraudulent Piltdown man fossil was initially accepted as an ancestor of humans precisely because it had a big brain. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Australopithecus_africanus_-_Cast_of_taung_child.jpg; https://commons.wikimedia.org/wiki/File:Piltdown_man.jpg).

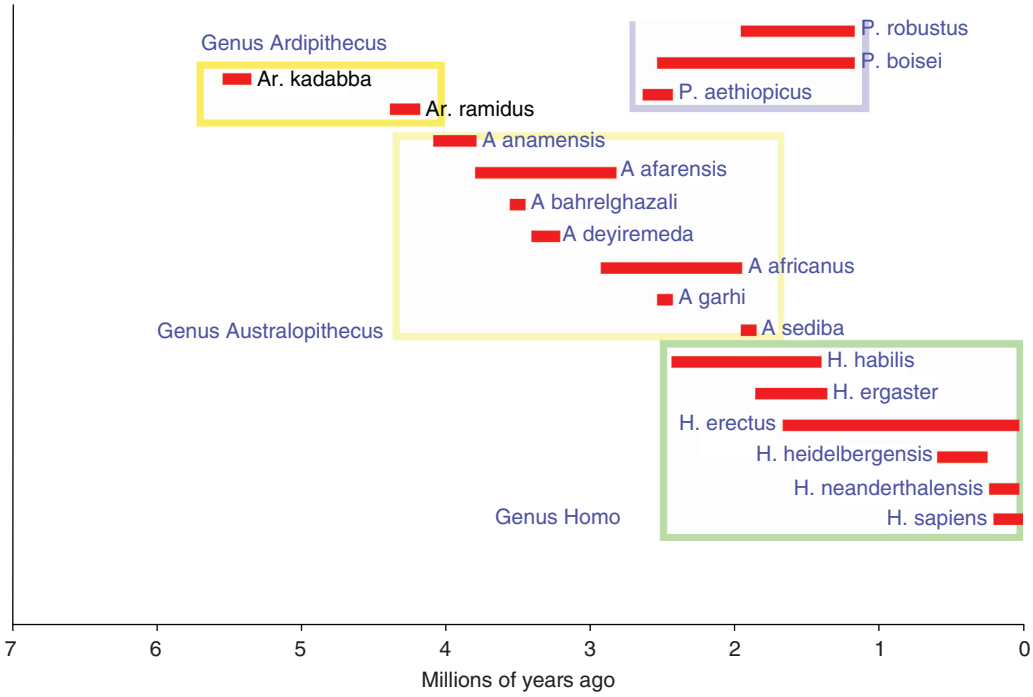


FIGURE 14.4

Current view of the distribution through time of all known hominin species. How these species might be related to one other, as well as which of these (if any) are ancestors of modern humans, remains a matter of much controversy. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Hominin_evolution.jpg).

The appearance of our genus, *Homo*, some 2–2.5 million years ago marks lots of changes; in addition to bigger brains, our ancestors are fully adapted to a bipedal lifestyle, show more and more evidence of group (social) behavior, and stone tools become increasingly more specialized. *Homo* also marks the first appearance of our ancestors outside of Africa, beginning around 1.5–2 million years ago—there is no credible evidence that australopithecines ever made it out of Africa. And once *Homo* gets out of Africa, they go everywhere, with *Homo* fossils found in Europe, East Asia, and even Indonesia by about 800,000 years ago or so. As with the australopithecines, there are many different named species of *Homo* (Figure 14.4), probably more than actually existed, but since you get a lot more kudos if you can claim you found the first example of a brand new species rather than yet another example of a previously described species, it's perhaps not surprising that anthropologists tend to want to interpret their new fossil finds as new species (and, of course, you get even more kudos if you claim that your new fossil is not only a new species but it overturns everything we thought we knew about human evolution!).

And what about our own species, *Homo sapiens*—what defines us as a species in the fossil record? It turns out that it is extraordinarily difficult to come up with a definitive description of what distinguishes our species, generally referred to as anatomically modern *Homo sapiens* (or AMHS for short), from other fossils ascribed to the genus *Homo*. Indeed, some attempts to come up with a list of defining characteristics for AMHS ran afoul of the problem that the definition would exclude some living populations of humans! Nevertheless, there are some general characteristics and trends that tend to set AMHS apart from our predecessors that can be identified, while recognizing that these are only general tendencies that may not be present in all AMHS (or absent in all fossils attributed to other species of *Homo*); see Figure 14.5 for a general overview of these trends.

How these various putative species of archaic *Homo* are related to one another as well as to us is a matter of heated debate. For example, based on strict criteria of falling within the range of variation of current human populations, the earliest fossils of AMHS occur roughly 100,000 years ago, and some would argue that this marks the beginning of our species; based on much looser criteria of having some features in common with at least some current human populations, others have argued that all *Homo* species within the past million years or so should be considered as part of our species.

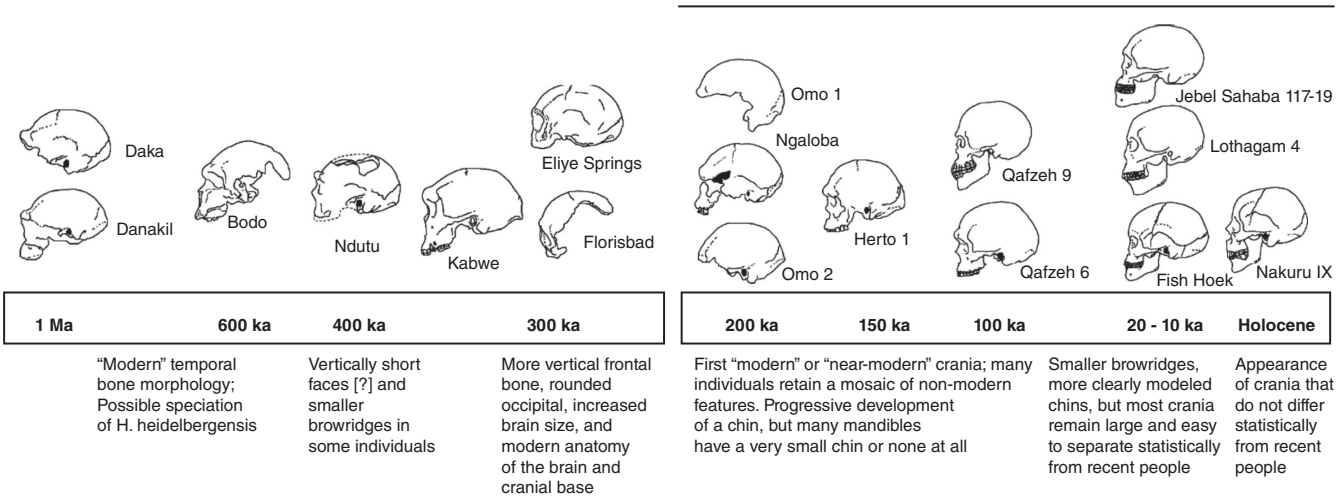


FIGURE 14.5

Changes in cranial features in the genus *Homo* associated with the transformation from archaic to modern humans. Reprinted with permission from Pearson, O.M., "Statistical and biological definitions of "anatomically modern" humans: Suggestions for a unified approach to modern morphology," *Evolutionary Anthropology* 17:38, 2008.

MODELS FOR HUMAN ORIGINS

Based on fossil evidence, there are four major models concerning human origins that have been put forward (Figure 14.6): the candelabra model; multiregional evolution; and a recent African origin (RAO) (which can be further divided into replacement vs. assimilation hypotheses). The candelabra model was the dominant model for many years—in fact, it was the model I was taught when I was a student—and gets its name from the fact that the branching pattern resembles a candelabra or candlestick. The basic premise of the candelabra model is that the ancestry of modern humans living in different parts of the Old World (viz., Africa, Europe, Asia, and Australasia) can be traced through a series of hypothetical ancestors in the fossil record back to a common origin sometime in the Miocene (upward of 2 million years ago or so). Clearly, this common ancestor of all Old World populations was not anatomically modern, since AMHS is somewhere between 0.1 and 1 million years old. It therefore follows that according to the candelabra model, the transition to AMHS must have happened independently, at four different times in four different parts of the Old World (or even more, since some anthropologists thought that there were additional transitions, e.g., two different transitions in Africa).

The candelabra model was most prominently associated with the anthropologist Carleton Coon, who was also known for his unabashedly racist views. In particular, Coon pointed out that since the candelabra model invokes independent origins of humans in different parts of the Old World, the transitions to AMHS need not have happened at the same time, or occurred precisely in the same way, in the different parts of the Old World (Coon 1962). Perhaps not surprisingly, since he was himself European, Coon argued that Europeans were the first to go through the transition to AMHS, and therefore have had the most time to evolve from their so-called "primitive" ancestry. And—surprise, surprise—according to Coon, Africans went through the transition to AMHS most recently, and therefore have had the least amount of time to evolve from their primitive ancestry.

We can see in the candelabra model in general, and in Coon's use of it to support his racist views in particular, the same sort of thinking about humans and how special we are alluded to in the beginning of the previous chapter. Namely, there is a tendency (consciously or subconsciously) to want to see evolutionary evidence that supports how special "we" are, whether "we" is defined as humans as opposed to other apes (as discussed in the previous chapter), or as Europeans as opposed to Africans. In the candelabra model, while a common ancestry of Europeans and Africans is (grudgingly) admitted, it is so distant in the past that Europeans and Africans became AMHS independently—it is hard to be more different, yet still within the same species, than that! Anyway, as anthropologists became increasingly uncomfortable with Coon's use of the candelabra model to promote his racist views, the candelabra model gradually fell out of favor. In hindsight, one wonders what took so long, as there are very good reasons for rejecting the candelabra model independently of how it might be used to promote racist views. Although we do not know for sure what all of the biological changes were that occurred in the transition to AMHS, they were undoubtedly so complex that to

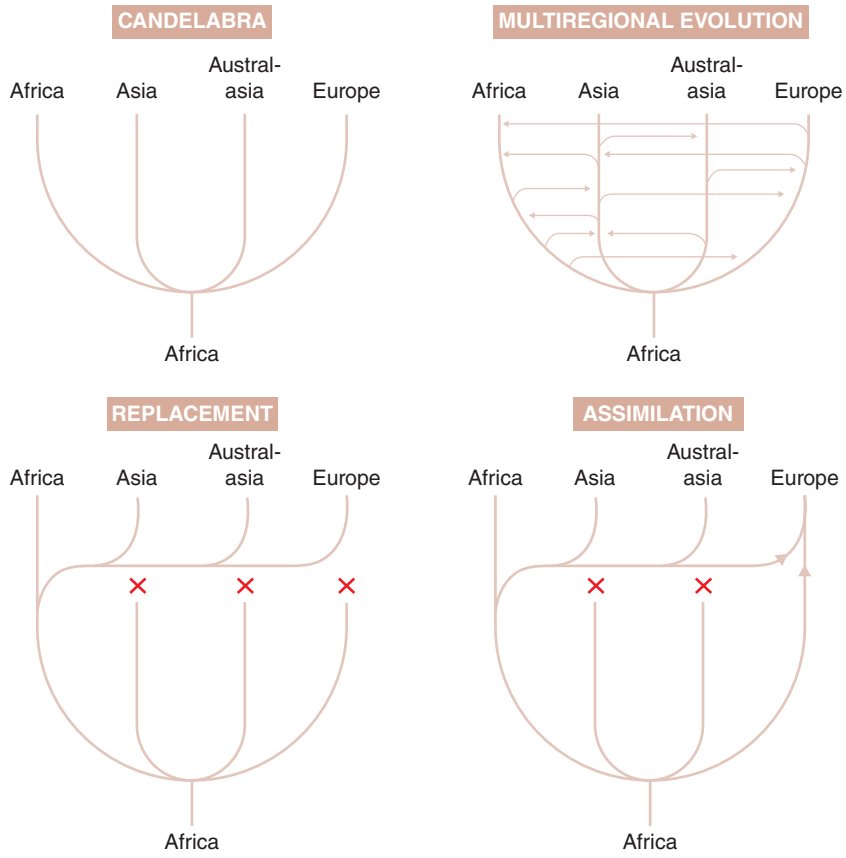


FIGURE 14.6

Four models for the origin of modern humans. The bottom two models reflect two versions of the Recent African Origin model. Reprinted with permission from Stoneking, M., "Human origins. The molecular perspective," *EMBO Reports* 9:S46, 2008.

propose they could have occurred completely independently four different times, in four different parts of the world, simply does not make sense in light of what we know about evolution. To be sure, there are some traits that seem to have evolved at least in part independently in different human populations—for example, lightening of skin pigmentation in Europeans and Asians, as will be discussed in more detail in the last chapter. But to suppose that AMHS as a species could have arisen completely independently (and multiple times at that) simply is not credible, and for that reason the candelabra model is not (to my knowledge) considered a reasonable explanation for our origins by any scientist. Still, when it comes to considering the genetic evidence for human origins, the candelabra model makes some predictions that contrast strongly with the other models, and so for that reason we will include the candelabra model in some discussions of the genetic evidence, while still stressing that nobody would argue in favor of this model.

So what came after the candelabra model? While there have always been many different ideas about human origins, beginning in the late 1970s and continuing for the next 20 years or so, there were two models that dominated the discussion. The first of these is multiregional evolution (Figure 14.6), which is most prominently associated with the anthropologist Milford Wolpoff and colleagues (Wolpoff et al. 1984), although a "polycentric" model, much like multiregional evolution, was proposed earlier by the anthropologist Franz Weidenreich (Weidenreich 1947). At first glance, multiregional evolution appears similar to the candelabra model, in that according to the multiregional evolution model, the main lines of ancestry are within each major region of the Old World (corresponding to the vertical lines in the model in Figure 14.6). That is, modern Europeans share features with ancient Europeans (such as Neandertals), modern Asians with ancient Asians, modern Africans with ancient Africans, and so forth—so, according to this model, there is regional continuity in the fossil record. At the same time, to avoid the problems associated with independent transitions to AMHS that plague the candelabra model, multiregional evolution also includes gene flow between populations (corresponding to the horizontal lines in Figure 14.6). So,

whenever a mutation arose that was important for the transition to AMHS, it was spread quickly via gene flow across the Old World regardless of where it arose, and thus the entire Old World population of hominins evolved in concert, over the last 1–2 million years, to become us.

The discerning reader may wonder how it is that regional continuity in certain fossil traits can be maintained in the face of the extensive gene flow needed to spread those mutations important for the transition to AMHS thousands and thousands of kilometers across the entire Old World, from sub-Saharan Africa to western Europe to eastern Asia. That is, why weren't the mutations responsible for those traits that supposedly link Europeans and Neandertals (for example) also spread across the Old World, thereby erasing the signal of regional continuity? The usual answer given by proponents of the multiregional evolution model is that selection would maintain those traits that show regional continuity and inhibit their spread elsewhere, because they have a selective advantage in that particular region of the Old World but not elsewhere. At the same time, selection would be operating to increase the frequency—all across the Old World—of those desirable mutations that led to us. So, multiregional evolution involves a rather complicated and delicate balance involving old mutations (that show regional continuity), recent mutations (for the transition to AMHS), gene flow to spread around these recent mutations, and selection to increase their frequency all across the Old World, as well as selection to maintain regional continuity. Whether or not it would all actually work was another source of debate—I can vividly remember meetings during the 1980s and 1990s during which population geneticists such as Masatoshi Nei would challenge proponents of multiregional evolution to provide specific details concerning the amounts of mutation, selection, and gene flow in the multiregional model, so these could be tested against the genetic evidence, but satisfactory answers were never forthcoming.

Anyway, the second major model of human origins was the RAO model (Figure 14.6), proposed by the anthropologists Stringer and Andrews (1988). In contrast to the multiregional evolution model, which holds that the entire Old World population of hominins evolved to become AMHS, the RAO model proposes that the transition to AMHS occurred within a single population. Other models had proposed a single origin for AMHS but placed the origin in Europe (of course!) or perhaps Asia; the RAO model differed in placing the origin in Africa, and moreover recently, beginning about 200,000 years ago or so, and ending with fully AMHS by 100,000 years ago or so. Then, beginning between 60,000 and 100,000 years ago, modern humans spread out of Africa, populating the entire Old World relatively quickly and ultimately spreading across the globe. Proponents of RAO pointed to features that early modern human fossils outside of Africa shared with earlier fossils within Africa as evidence for the model.

Beginning in the mid-1980s, a fierce debate raged at conferences and in scientific publications between the proponents of multiregional evolution and those favoring the RAO model. Arguments centered over whether or not particular fossils exhibited key features that would support regional continuity versus other key features that would support links to Africa. Proponents of the RAO model also suggested other explanations for regional continuity, such as retention of ancestral polymorphism (Figure 14.7). There were also alternative versions of the RAO proposed, with some arguing for complete replacement of all non-African archaic humans, while others proposed some interbreeding between modern humans coming from African and at least some archaic humans outside of Africa (e.g., Bräuer 1989); these are shown as the "Replacement" and "Assimilation" models in Figure 14.6.

Why was it so difficult to distinguish between these models on the basis of fossil evidence? I think it is at least partly because of the subjectivity inherent in such analyses—e.g., we don't have living, breathing Neandertals or other archaic hominins around to compare

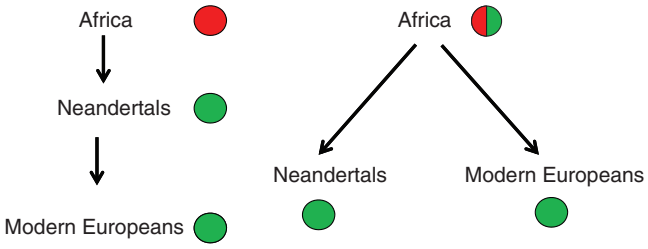


FIGURE 14.7

Regional continuity versus ancestral polymorphism as competing explanations for sharing of particular features between archaic and modern humans. In the diagram below, the red and green circles depict two states for the same feature (i.e. presence/absence). Under regional continuity, the red feature characterizes ancestral Africans, the green feature arises in the archaic humans outside Africa (here represented by Neandertals), and then also shows up in modern humans because of inheritance from Neandertals. Under ancestral polymorphism, both the red and green features are present in Africans, who give rise to archaic humans and then later to modern humans; the archaic humans and modern humans share the green feature either by chance or because of a possible selective advantage for the green feature outside Africa.

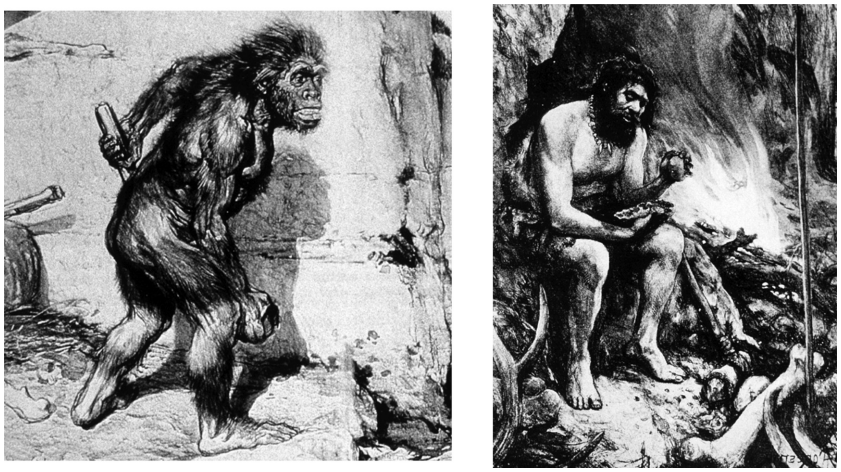


FIGURE 14.8

Two reconstructions based on the same Neandertal fossil, from La Chapelle-aux-Saints. Left, by František Kupka, based on the work of Marcellin Boule, from the Illustrated London News, 1909. Right, by Amédée Forestier, based on the work of Arthur Keith, from the Illustrated London News, 1911, reprinted with permission.

ourselves to, so instead we must make use of reconstructions based on bits and pieces of their fossils. And (consciously or subconsciously) biases can creep into these reconstructions. As just one example of what can happen, see Figure 14.8, which shows two different reconstructions based on the same famous (and rather complete) Neandertal skeleton from La Chapelle-aux-Saints, France. On the left, you see a hairy, brutish, fierce-looking creature carrying a big club, no doubt ready to bash in the skull of some poor unsuspecting modern human who is coming around the corner. And you look at that reconstruction and you think, well, if that's what Neandertals were like, maybe they were *your* ancestors, but they certainly weren't *my* ancestors. But on the right, you see someone who looks very much like a modern human, wearing a necklace (which is an artistic stretch since Neandertals did not appear to have such jewelry, but never mind), making a stone tool, and obviously thinking very hard about what he is doing. And you look at that and you think, OK, if that's what Neandertals were like, then I have no problem with them being my ancestors. The truth, of course, is that neither of these are very good reconstructions, and rather than telling us anything about Neandertals, they instead tell us what the people who made these reconstructions thought Neandertals were like.

So, even though the multiregional evolution and RAO models were based on fossil evidence, it was extraordinarily difficult to distinguish between them from the fossil evidence. Actually, it turns out that the best way to distinguish between these models (in my opinion, anyway) is not the fossil evidence but rather genetic evidence, because these different models of human origins are really statements about genes. As shown in Figure 14.9, the different models of human origins make different predictions concerning the amount of African ancestry we should expect to find in non-African populations. At one extreme is the candelabra model (again, included here for purely illustrative purposes, not because anyone takes it seriously). According to the candelabra model, modern Europeans got all of their genetic ancestry from ancient Europeans, modern Asians got all of theirs from ancient Asians, and so forth. Therefore, the candelabra model predicts 0% African ancestry in non-African populations.

At the other extreme is the RAO model with complete replacement (called, for simplicity, the replacement model). In this version of the RAO model, after modern humans spread out of Africa into Europe and Asia, they completely replaced the non-African archaic hominins without any interbreeding. Therefore,

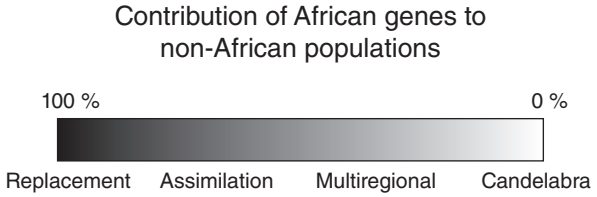


FIGURE 14.9

Predictions of the different models of human origins concerning the amount of African ancestry expected in non-African populations. Reprinted with permission from Stoneking, M., "Mitochondrial DNA variation and human evolution," in Human Genome Evolution, M. Jackson, T. Strachan, and G. Dover (editors), BIOS Scientific Publishers: Oxford, pp. 263–281, 1996.

modern non-Africans trace all of their ancestry to the RAO, and hence there should be 100% African ancestry in non-African populations.

And in between these two extremes, there would be the multiregional evolution and assimilation models. According to multiregional evolution, while there would be some African genes in non-African populations, the major part of the ancestry in European populations would be from Europe, in Asian populations from Asia, and so forth (to account for regional continuity); the multiregional model is quite firm in rejecting any massive movement of people (and their genes) out of Africa (Wolpoff et al. 1994). Whereas, according to the assimilation model, while the major fraction of the ancestry of non-African populations would be African (to reflect the RAO), some (small) proportion of the ancestry would be non-African in at least some populations (e.g., Europeans might have a small contribution of Neandertal genes if interbreeding took place with Neandertals).

Predictions about the overall age of genetic variation within humans also differ for these models. To be sure, because ancestral populations will be polymorphic, the ages of mutations will often be older than the populations themselves, as discussed previously in Chapter 12. Still, by proposing that AMHS arose from a single population in Africa, the RAO model would suggest an overall much more recent age for the genetic variation in AMHS than multiregional evolution, which proposes that the entire Old World population of archaic humans contributed genetically to AMHS since the first exodus of our ancestors from Africa, around 2 million years ago. In sum, to distinguish among these various models of human origins, we should turn to the genetic evidence.

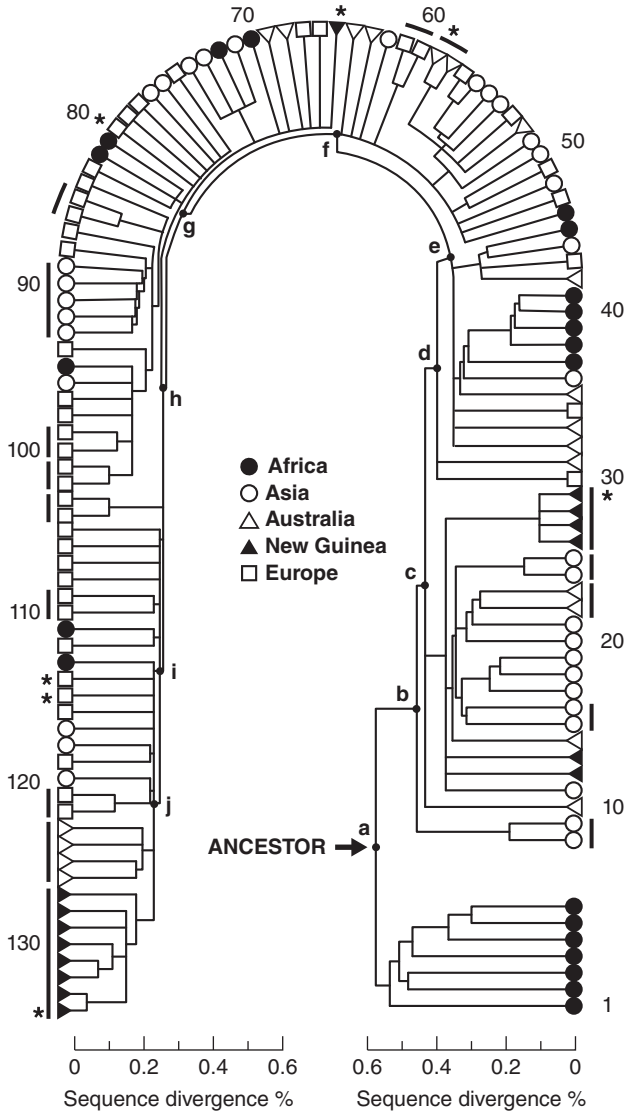


FIGURE 14.10

Maximum parsimony tree of human mtDNA types. Black bars indicate clusters of mtDNA types from the same geographic region; asterisks indicated mtDNA types found in more than one individual. Reprinted with permission from Cann, R., et al., "Mitochondrial DNA and human evolution," *Nature* 325:31, 1987.

THE GENETIC EVIDENCE: mtDNA

The first DNA evidence to address the issue of human origins in detail came from studies of human mtDNA variation in the laboratory of Allan Wilson (of Sarich and Wilson fame) that I participated in when I was a graduate student. Working with RFLP variation in mtDNAs purified from a worldwide sample of placentas, Rebecca Cann (another graduate student) and I found that Africans harbored the most mtDNA variation—about twice as much, on average, as Europeans—and that the mtDNA variation outside of Africa appeared to be a subset of the variation in Africa (Cann et al. 1987). As discussed in Chapter 12, this result strongly suggests an African origin of human mtDNA variation. A phylogenetic (maximum parsimony) tree relating the different mtDNA types (Figure 14.10) also found support for an African origin, as the root of the tree (obtained by midpoint rooting, so assuming a molecular clock) divided it into two primary branches, one consisting of only African mtDNA types, the other consisting of all of the non-African mtDNA types as well as some African mtDNA types. The exclusive presence of African mtDNA types on both primary branches of the tree is another indication of an African origin. Finally, a molecular clock approach dated the origin of all of the human mtDNA diversity to about 200,000 years ago; all of these results support the predictions of the RAO model.

When these results were published in 1987, they were dismissed by some as too ridiculous to merit serious consideration—just as Wilson's earlier work with Sarich on human–ape relationships was treated. Moreover, there was some outright misunderstanding as to just what the concept of a common maternal ancestor of all human mtDNAs actually meant, with some taking it to mean that all humans (and all of our genes) were descended from a single African female, rather than all human mtDNAs (go back and reread the discussion on this point in Chapter 12, especially concerning Figure 12.5, if you are similarly confused). It didn't help matters any when journalists started calling the mtDNA ancestor "Eve," and newspaper and magazine articles appeared saying that yes, there was an Eve, and she was black!

However, there was also much legitimate criticism and discussion. In particular, attention focused on the use of African–Americans as the primary source of African mtDNAs, because the known European admixture in African–Americans might elevate their levels of genetic variation above that of native Africans. We argued that most of the European admixture into African–Americans involved European–American males and African–American females, as a consequence of the American slavery practices, and hence African–Americans should have mostly African mtDNA types. Subsequent studies have confirmed this assertion (Parra et al. 1998). Critics also questioned the molecular clock approach, and the accuracy of phylogenetic reconstruction of such large data sets—this was one of the first times that maximum parsimony had been carried out on a data set of this size.

Subsequent studies tried to address these issues by obtaining more data and developing better methods of analysis, and during the 1990s, the pendulum swung back and forth between studies that seemed to show more mtDNA support for the RAO model (e.g., Vigilant et al. 1991) and those that seemed to call such results into question (e.g., Hedges et al. 1992). However, any issues regarding the mtDNA evidence have been convincingly laid to rest; Figure 14.11 shows a tree of human mtDNA types from a global sample that is based on complete mtDNA genome sequences. Rooted by using a chimpanzee mtDNA genome sequence as an outgroup, the tree divides into two primary branches, with only mtDNAs from Africa on both primary branches, and an estimated date for the mtDNA ancestor of about 150,000 years ago—overall, remarkably similar to the results we published in 1987! This study, as well as numerous subsequent studies, provides overwhelming support for an RAO of human mtDNA. And as we shall see later, this conclusion is further supported by mtDNA sequences from Neandertals. However, it must be kept in mind that mtDNA is

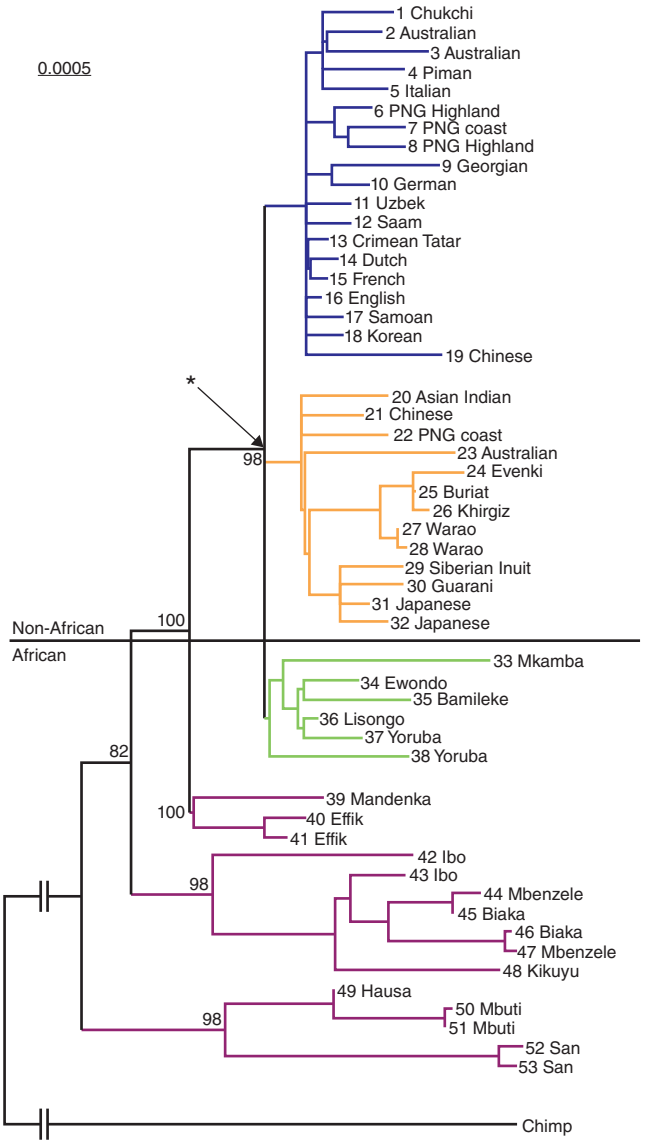


FIGURE 14.11

Phylogenetic tree of complete mtDNA genome sequences, rooted with a chimpanzee mtDNA sequence. The arrow and asterisk point to a part of the phylogeny that shows evidence for population expansion, associated primarily (but not exclusively) with non-Africans. Numbers on branches are bootstrap support values. Reprinted with permission from Ingman, M., et al., "Mitochondrial genome evolution and the origin of modern humans," *Nature* 408:708, 2000.

just a single genetic locus, and the history of a single gene can differ from that of a population or species, either because of chance effects (i.e., genetic drift) or because of selection on that gene. So, mtDNA alone is not enough to allow us to say that the RAO model explains human origins; for that purpose, we must look at additional genes and see what they have to say.

THE GENETIC EVIDENCE: Y CHROMOSOME

As discussed previously, the male counterpart to the maternally inherited mtDNA is the Y chromosome, so it is natural to ask how the NRY (nonrecombining part of the Y chromosome) story compares to the mtDNA story. And, as recounted earlier, studies of NRY variation lagged behind those of mtDNA due to the lack of suitably variable genetic markers on the NRY. But in a landmark study, Luca Cavalli-Sforza and colleagues published the first in-depth survey of human NRY variation (Underhill et al. 2000). And the results (Figure 14.12) are remarkably similar to the mtDNA results, in that the tree of NRY types from a worldwide sample of humans strongly indicates an African origin of human NRY variation, as the deepest splits within the tree involve exclusively African lineages. However, dating the origin of our NRY ancestor has proven to be more elusive. Because of the way the polymorphic markers on the NRY were ascertained, there is no simple way to figure out how fast they are evolving. Initial attempts to date the age of the NRY ancestor, therefore, utilized STR (short tandem repeats; see Chapter 7 if you need a refresher on what they are) markers or limited amounts of sequence data and came up with dates of around 60,000–100,000 years ago (Pritchard et al. 1999; Thomson et al. 2000), so more recent than the mtDNA ancestor. This younger TMRCA for the NRY was generally attributed to a smaller effective population size for males than for females, which could reflect lower rates of male migration (discussed in Chapter 19) or fewer males than females participating in reproduction (Wilkins 2006). Next generation sequencing has changed all that, with the most recent estimate for the age of the ancestor of the phylogeny shown in Figure 14.12 (and schematically in Figure 9.7) clocking in at between 160.000 and 260,000 years ago, depending on what mutation rate is used (Barbieri et al. 2016).

However, an additional complication to the Y chromosome story is the discovery of an extremely divergent lineage called A00 (Mendez et al. 2013), which roots outside the previously known phylogeny and

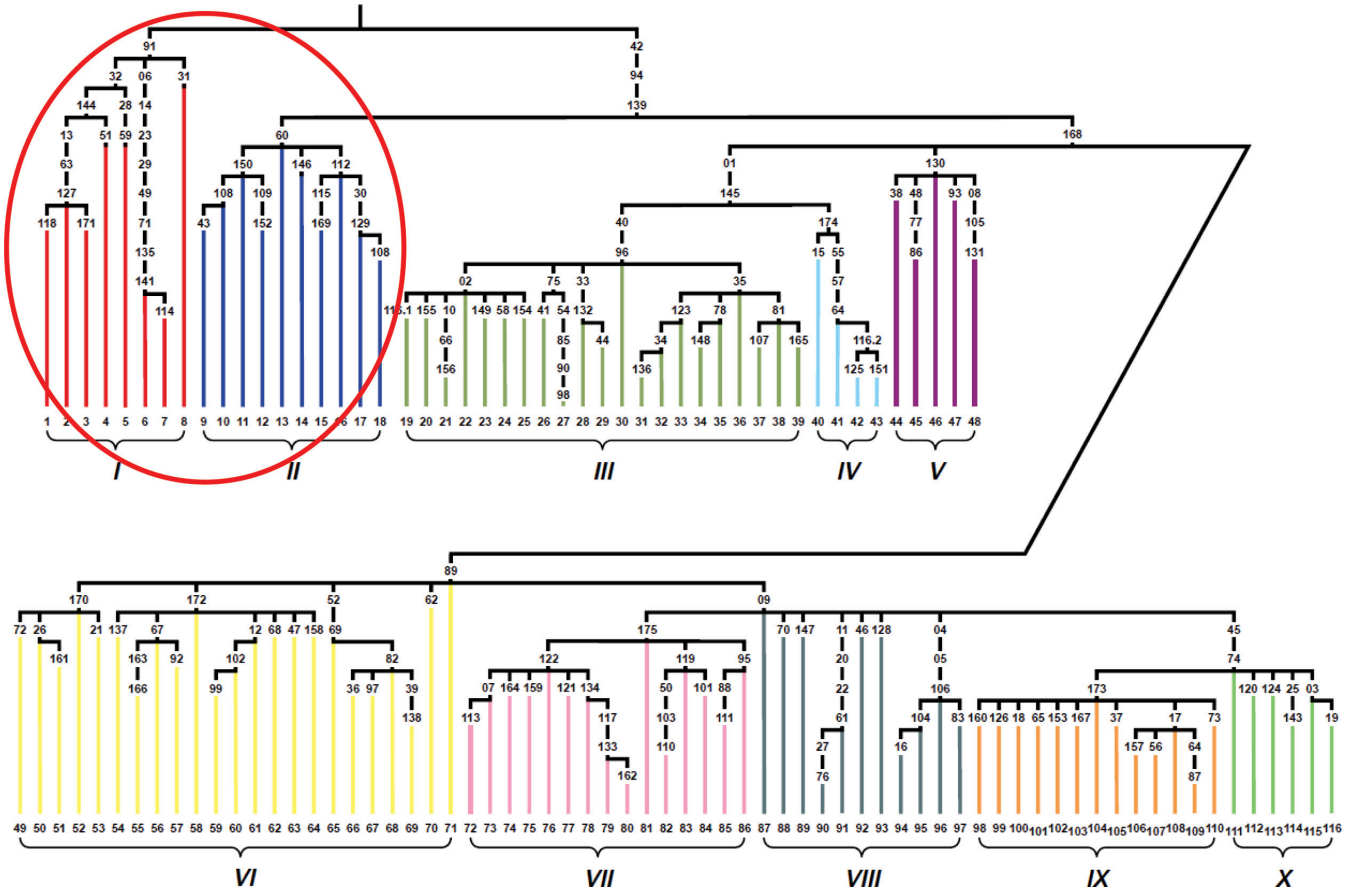


FIGURE 14.12

Phylogeny of Y chromosome types. The first two branches (circled) are found exclusively in Africans and hence provide support for an African origin of modern human Y chromosome diversity. Modified with permission from Underhill, P., et al., "Y chromosome sequence variation and the history of human populations," *Nature Genetics* 26:358, 2000.

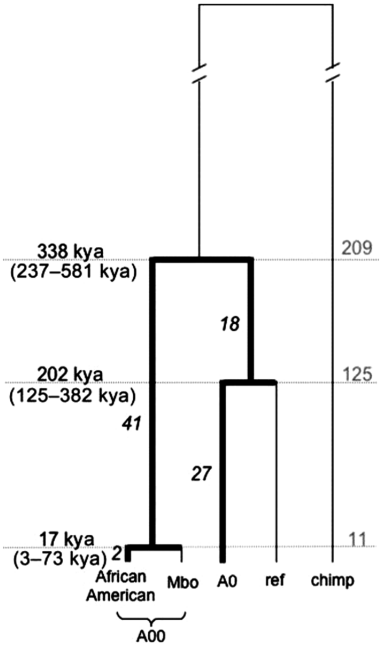


FIGURE 14.13

Phylogenetic tree showing the relationship of the A00 lineage, found in an African–American and subsequently in several individuals from the Mbo group of Cameroon, to other human Y chromosome lineages (A0 and ref, for reference sequence) and to the chimpanzee Y chromosome sequence. Numbers on branches indicate mutational differences; black numbers are ages obtained using a slower rate of Y chromosome evolution and gray numbers are ages obtained using a faster rate of Y chromosome evolution. Reprinted with permission from Mendez, F.L., et al., "An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree," *American Journal of Human Genetics* 92:454, 2013.

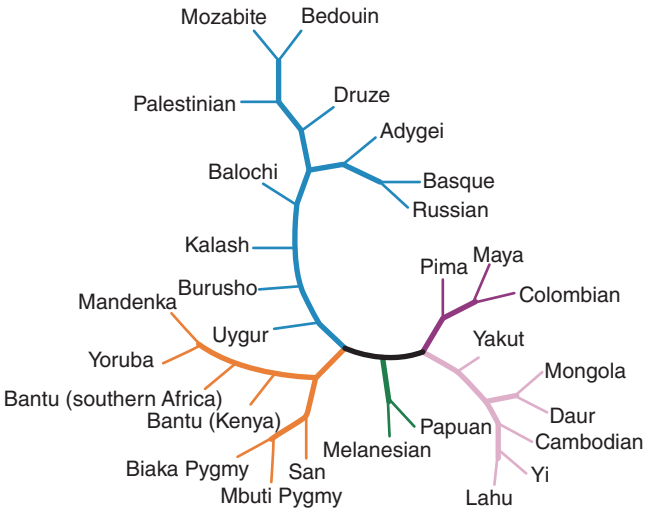


FIGURE 14.14

Neighbor-joining tree for 29 populations from the CEPH–HGDP panel, genotyped for about 526,000 SNPs. Bootstrap values are not shown because they are all quite high (more than 95%). Reprinted with permission from Jakobsson, M., et al., "Genotype, haplotype and copy-number variation in worldwide human populations," *Nature* 451:998, 2008.

diverged some 209,000–338,000 years ago, again depending on the mutation rate (Figure 14.13). This lineage was first discovered in an African-American (through one of the personal genomics companies—a nice example of how personal genomics can contribute to science) and subsequently in some individuals from Cameroon. The much older age of this lineage could reflect introgression from an archaic human, or it could also reflect deep population structure within modern humans in Africa; the jury is still out on this one.

THE GENETIC EVIDENCE: AUTOSOMES

Along with the mtDNA and NRY studies, beginning in the late 1980s increasing attention was paid to analyses of autosomal DNA variation that could address the dispute between multiregional evolution and RAO concerning human origins. Various types of markers (RFLPs, STRs, Alu insertions, etc.) were analyzed in various worldwide samples of human populations. And typically (albeit by no means unanimously), the general type of result that was obtained seemed to favor RAO over multiregional evolution, in that African populations showed the greatest divergence and/or the deepest splits in tree analyses, with relatively recent dates for the age of the diversity.

More recently, genome-wide data (from either SNP chips or full sequences) have both confirmed and extended this picture. For example, Figure 14.14 shows a tree of population relationships, based on ~650,000 SNPs, that shows the same results that we've already seen time and time again that are indicative of—you guessed it—an African origin. But we can do a lot more with genome-wide data than simply build the same old boring trees over and over again. For example, we can use the methods of demographic inference discussed in Chapter 12 to estimate various parameters (divergence time, effective population size, etc.) concerning the history of human populations from such genome-wide data, and an example is shown in Figure 14.15. As this figure shows, genomic data strongly support the occurrence of the deepest population splits within Africa, divergence between African and non-African populations around 75,000–100,000 years ago, a bottleneck (population decrease) associated with the

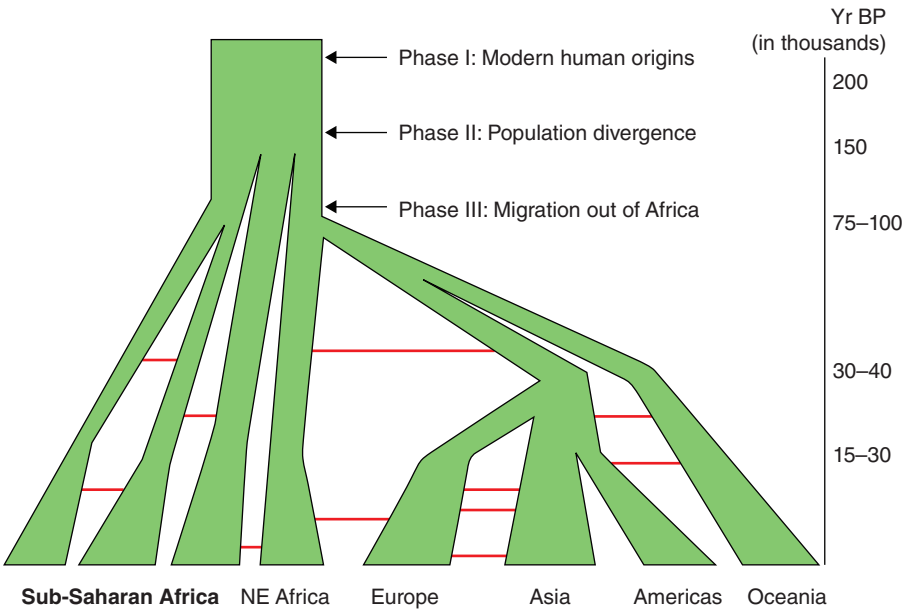


FIGURE 14.15

Demographic history of human populations. Widths of the green branches are proportional to effective population size, while red lines indicate gene flow (in a general way, not specific migrations). Reprinted with permission from Tishkoff, S.A., and Williams, S.M., "Genetic analysis of African populations: human evolution and complex disease," *Nature Reviews Genetics* 3:611, 2002.

expansion out of Africa, and more recent population increases worldwide.

With full genome sequences, we can also infer the history of population size changes over time, using the PSMC method described in Chapter 12 (with Figure 12.19 illustrating the method); an example is shown in Figure 14.16. There is remarkably good congruence in the PSMC curves for the different genome sequences from about 2 million until about 100,000 years ago, which makes sense as human populations began diverging only around 100,000 years ago—prior to this time, all populations should have the same history. The curves show that our population size was decreasing until about

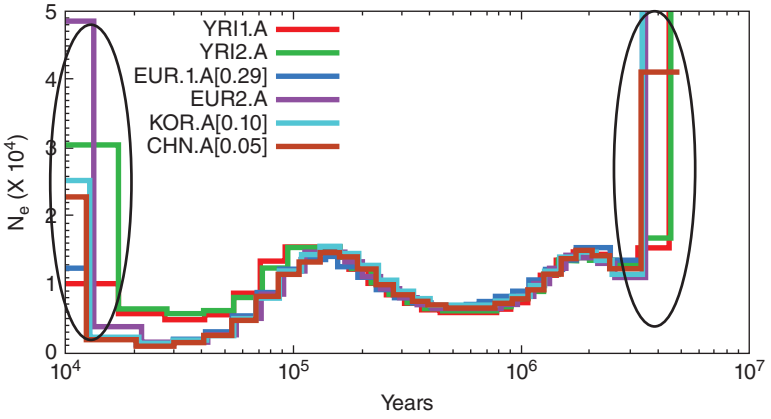


FIGURE 14.16

PSMC curves of population size change over time derived from full genome sequences from two Yoruba (YRI), two Europeans (EUR), a Korean (KOR), and a Chinese (CHN). The circled areas indicate time periods where the PSMC is not informative, as there are not enough genome segments that date to those times to provide a good estimate of the effective population size. Note that time is shown as years ago, proceeding from younger to older from left to right, and is plotted on a logarithmic scale. Reprinted with permission from Li, H., and Durbin, R., "Inference of human population history from individual whole-genome sequences," *Nature* 475:493, 2011.

400,000–500,000 years ago and than began increasing until about 100,000 years ago. At this point, human populations begin diverging—presumably corresponding to the out-of-Africa migration—and also begin decreasing in size, with a bigger decrease in non-African than African populations (presumably reflecting the out-of-Africa bottleneck), although it is interesting that African populations also show a population decline. Population size then begins increasing around 10,000–20,000 years ago, but this is the point at which the PSMC approach is no longer informative, as there are too few segments in the genome with a TMRCA this recent to get accurate estimates of the effective population size. There is clearly a lot of information in our genome about our past history, and inferring demographic history from genome-wide data is a very active area of current research, with new approaches appearing all the time.

And there are other aspects of the data we can look at. For example, Franck Prugnolle and colleagues examined the relationship between how much genetic diversity a population has and how far it is from East Africa (Prugnolle et al. 2005). They reasoned that if modern humans did indeed arise in Africa and then spread out of Africa across the Old World, then genetic diversity might be correlated with distance from Africa. In calculating how far a population was from Africa, they did not use straight-line distances from a map but instead took into account how humans were likely to have migrated, for example, going around mountains or major bodies of water rather than directly across them. The result, shown in Figure 14.17, indicates an extraordinarily close relationship between the genetic diversity of a population and how far it is from Africa that is remarkable for two reasons. First, it suggests that if you want to know how much genetic diversity a population has, you don't need to go to all the effort of obtaining research permits and mounting an expedition to collect samples, followed by laborious laboratory work—simply figure out how far the population is from East Africa, then use the graph shown in Figure 14.17 to estimate the genetic diversity! Second (and more seriously), the obvious explanation for the relationship for the graph in Figure 14.17 is that modern humans originated in Africa, left Africa via East Africa, and spread across the world via a series of successive bottlenecks (decreases in population size), accounting for the decreasing genetic diversity as one gets further and further from Africa. This is an example of **serial bottlenecks**, discussed back in Chapter 12, and in fact it is really hard to come up with any other explanation for this graph.

The overwhelming conclusion from the multilocus autosomal studies is thus like that from mtDNA and the NRY; namely, there is a genome-wide signal of an RAO for AMHS. But while these results may be sufficient

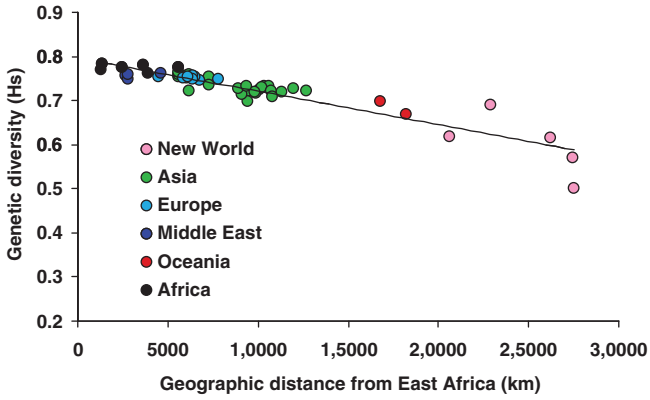


FIGURE 14.17

Plot of genetic diversity (heterozygosity, Hs) versus geographic distance from Africa (using migration routes that go around rather than through major geographic obstacles to human migration). Genetic diversity estimates are from the CEPH–HGDP populations and are based on 377 STR loci. The overall decrease in genetic diversity with increasing distance from East Africa strongly supports an African source for non-African genetic diversity via serial bottlenecks. Reprinted with permission from Prugnolle, F., et al., "Geography predicts neutral genetic diversity of human populations," *Current Biology* 15:R159, 2005.

to reject multiregional evolution (not to mention the candelabra model, already rejected on other grounds) in favor of the RAO model, what about the replacement versus assimilation versions of the RAO model? Here the results are more equivocal, as the genome-wide analyses tend to pick out the dominant signals from across the genome and may miss subtle signals of a genetic contribution from non-African archaic humans that would favor assimilation over replacement. In-depth studies of single autosomal loci or genomic regions (similar to the mtDNA and NRY studies) are needed to see whether there is any signal in our genomes that would suggest assimilation rather than replacement.

What would such a signal look like? One approach is that if an RAO with complete replacement indeed holds, then any mutations that arose outside of Africa have to be younger than the exodus of AMHS from Africa, which would place an upper bound for such mutations of at most 200,000 years ago (allowing for variance in the molecular clock estimates). Any mutation older than this must have occurred in African ancestors of AMHS, if the replacement version of RAO is correct. So, if we can find any mutations that seem to have arisen outside of Africa (because they are not found in African populations) that are more than 200,000 years old, the inference is that such mutations must have arisen in a non-African archaic

population and were contributed to AMHS via interbreeding between the archaics and AMHS—in other words, assimilation.

Over the past 15 years or so, numerous in-depth studies of single genetic loci or genomic regions were carried out. While most of them did indeed find patterns of variation consistent with an RAO—that is, most diversity in Africa, deepest lineages within Africa, and so forth—a few did not. So, do these studies provide unequivocal support for the assimilation version of the RAO hypothesis? Unfortunately, it's not as simple as that, as there are other explanations for an older-than-expected non-African mutation. First, the mutation may indeed exist in Africa, and therefore may have spread from Africa, but has not been sampled yet in Africa. This in fact turned out to be the case for some of these claimed cases, as frequently African populations are underrepresented in human genetic diversity studies; the mutation in question then turned up in Africa when more African populations were sampled. Second, selection can distort patterns of variation and cause the variation outside of Africa to appear older than it really is. Third, even in the absence of selection, the mutation may have arisen in Africa and spread out of Africa with migrations of AMHS but was subsequently lost in Africa via genetic drift. Simulation studies have shown that this can indeed happen at low but not insignificant frequencies, even with complete replacement, and therefore such apparently old non-African mutations are not conclusive proof of interbreeding and assimilation.

Thus, beginning with the early mtDNA studies of the 1980s and continuing until just a few years ago, the general consensus waxed and waned as to whether the genetic data supported replacement or assimilation. No real resolution was forthcoming from genetic analyses of current populations, as for every argument or data set put forward in favor of assimilation, a counterargument demonstrating that the finding in question was also consistent with replacement was put forward (or vice versa). However, the argument over replacement versus assimilation has now been resolved (at least, to most people's satisfaction—when it comes to human origins, the only thing one can say with 100% certainty is that there will never be 100% agreement on anything!), and the resolution came about not from genetic analyses of current populations but rather from ancient DNA analyses, which we will turn to in the next chapter.

CHAPTER 15

ANCIENT DNA

So far, we have restricted ourselves to analyses and results coming from studies of contemporary populations. However, there is another source of information about our genetic history, and that is ancient DNA, which is DNA extracted from fossils or other ancient materials, and that is the topic of this chapter. You may think, well, what's the big deal, DNA is DNA, and you would be partly right—in many respects, ancient DNA is analyzed just like contemporary DNA. But there are some very important aspects in which ancient DNA analysis differs from the analysis of contemporary DNA, so before discussing what we have learned from ancient DNA, we need to first appreciate the different properties of ancient DNA.

PROPERTIES OF ANCIENT DNA: DEGRADATION

Immediately following the death of an individual, DNA starts degrading. Nucleases (enzymes that degrade DNA) and other chemicals in your body that are normally kept separate from DNA gain access to DNA as cell walls and nuclear membranes start disintegrating. These nucleases break the DNA into smaller and smaller fragments. Environmental sources also contribute to this process of DNA degradation—water, heat, ultraviolet rays from the sun, and so forth. The end result is that DNA in ancient remains generally exists in much lower amounts, and in much smaller fragment sizes, than the DNA you would get from a "fresh" sample. This, in turn, requires special procedures for isolating the DNA from an ancient specimen—it is all too easy to lose the DNA during the extraction process when there isn't much of it to begin with—as well as special procedures for analyzing small fragment sizes. A PCR assay to amplify a 1000-bp size fragment of DNA may work quite well with fresh DNA but would in all likelihood fail miserably with ancient DNA, simply because there are no fragments of that size remaining in the DNA from the ancient specimen.

PROPERTIES OF ANCIENT DNA: DAMAGE

In addition to decreasing the total amount of DNA and the size of DNA fragments, ancient DNA is also frequently chemically modified. Recall from Chapter 2 that our DNA is constantly under attack and modification by intracellular by-products of metabolism (such as free radicals) as well as the environment (such as ultraviolet light), and we accordingly have evolved elaborate and sophisticated mechanisms to recognize and repair such DNA damage; without such repair mechanisms life would not be possible. After death, of course, all DNA repair ceases, so damage then accumulates in the DNA. This damage takes a variety of forms, such as nicks between bases in a strand of the DNA helix, double-strand breaks, cross-links between DNA strands, loss of bases, and various modifications of bases due to oxidation, depurination, and so forth (the latter are shown in Figure 15.1).

These chemical modifications have two potential consequences, depending on the type of modification. The first is that the damaged DNA may be completely refractory to the analysis—cross-linked DNA, for example, cannot be amplified by the DNA polymerases typically employed in PCR or other analyses; it's as if the DNA simply isn't even present. The second possibility is that a modified base is mistakenly recognized as a different base. By far the most common type of damage that occurs to single bases in DNA is deamination of cytosine to uracil (Figure 15.2), which occurs spontaneously in the presence of water. Recall that uracil is one of the four bases found in RNA but is not a normal component of DNA—uracil (U) in RNA takes the place of thymine (T) in DNA. When DNA containing a uracil is amplified with a DNA polymerase (e.g., during PCR or the preparation of sequencing libraries for next-generation sequencing), the DNA polymerase acts as if the uracil is really a thymine and inserts an adenine on the opposite strand. The end result is that a CG base-pair in the original DNA is converted to a TA base-pair in the amplified DNA. While other bases

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

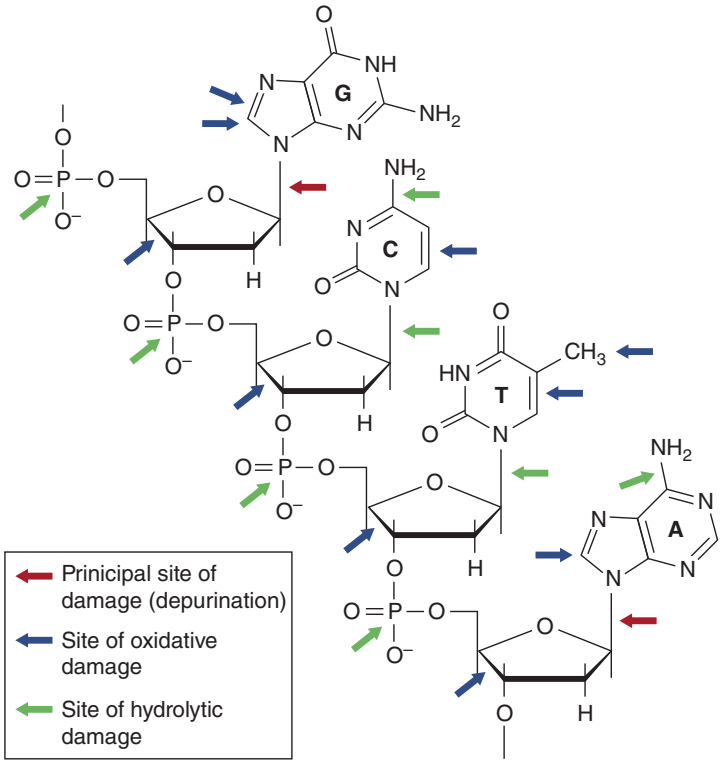


FIGURE 15.1

Different types of DNA damage that can occur after death, including loss of purines (recall that A and G are purines while C and T are pyrimidines), damage due to exposure to oxygen (oxidative damage), and damage due to exposure to water (hydrolytic damage). Reprinted with permission from Hofreiter, M., et al., "Ancient DNA," *Nature Reviews Genetics* 2:353, 2001.

can also undergo deamination or other modifications, it turns out that cytosine deamination is by far the most important source of damage in ancient DNA. This can be readily seen in clones of PCR products from ancient DNA, as well as in next-generation sequences from ancient DNA (Figure 15.3).

One way to deal with this issue of cytosine deamination is to treat the ancient DNA extract with an enzyme called uracil *N*-glycosylase (UNG, also known as uracil deglycosylase or UDG), which is part of the normal cell machinery for getting rid of uracil in DNA (which can occur by cytosine deamination or by mistaken incorporation of uracil in place of thymine during DNA replication). Uracil *N*-glycosylase treatment excises uracil from DNA, leaving behind an abasic site (i.e., a position in one strand of the DNA without a base, so the base on the opposite strand is unpaired). In living cells, the abasic site would then be repaired, using the unpaired base on the opposite strand to insert the correct base. However, in an ancient DNA extract no such repair can occur, so the abasic sites remain and lead to strand breaks in the DNA that prevent PCR or other amplification. The result is that UNG treatment enriches ancient DNA extracts for fragments that lack cytosine deamination, which is both a blessing and a curse. It is a blessing because there is a much lower chance of mistaking postmortem cytosine deamination for a true mutation, but it is a curse because the already limited amount of DNA in an ancient DNA extract is further reduced. In fact, if there is a lot of cytosine deamination, there may not be any undamaged fragments left for analysis. Moreover, the signature of cytosine deamination evident in cloned sequences from PCR products or in next-generation sequences (Figure 15.3) turns out to provide useful confirmation that the sequences in question really do derive from ancient DNA and not from contamination of the extract with modern DNA (as discussed in more detail in the "Properties of ancient DNA: contamination" section). So, while UNG treatment can be useful in certain circumstances, it is often better to simply leave the extracts alone and deal with the cytosine deamination in the data analysis.

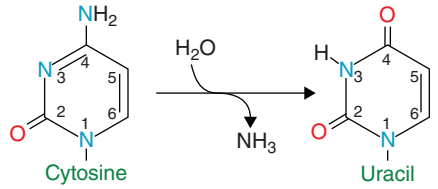


FIGURE 15.2

Hydrolytic deamination of cytosine, the most common form of DNA damage observed in ancient DNA. In the presence of water, cytosine loses an ammonia (NH_3) group and is converted to uracil, which is interpreted as a T by the DNA polymerase used in PCR and other manipulations of ancient DNA. Thus, ancient DNA typically shows an excess of C → T changes due to this hydrolytic deamination.

```
REF     ACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCCACCCCTCACCCAC
Clone1  ..............C...........T...........................T...T....
Clone2  .T............C.....................................................
Clone3  ..............C.....................................T..T........
Clone4  ..............C.....................................................
Clone5  ..............C.....................................................
Clone6  ..............C.....................................................
Clone7  ..............C.....................................T...........
Clone8  ..............C.....................................................
```

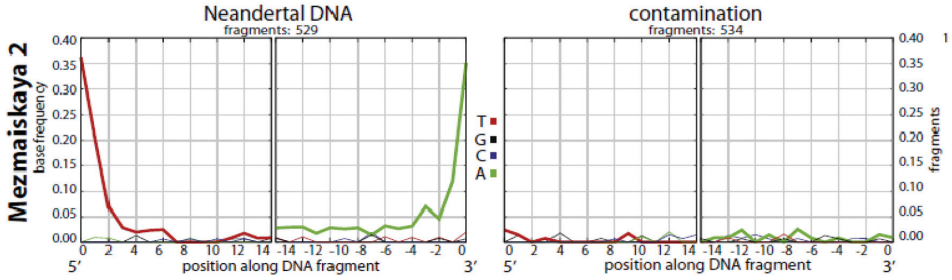


FIGURE 15.3

Top, sequences of several clones from Ötzi the Iceman, compared to the reference sequence, for a portion of the mtDNA genome. Identity to the reference sequence is indicated by a dot; differences are indicated by letters, with the blue letter indicating a difference observed in all clones that is probably a true difference from the reference sequence, and the white letters indicating sporadic differences indicative of damage (especially because these are all C → T changes). Bottom, excess of C-T and G-A changes at the ends of authentic ancient DNA reads (but not contaminating modern DNA reads) from next-generation sequencing of Neandertal mtDNA. The ends of ancient DNA fragments tend to be single-stranded, and cytosine deamination occurs more rapidly in single-stranded DNA. Because of how DNA sequencing libraries are prepared, this results in an excess of C-T changes in the forward direction and an excess of G-A changes in the reverse direction. Top, data from Handt, O., et al., "Molecular genetic analyses of the Tyrolean Ice Man," *Science* 264:1775, 1994. Bottom, reprinted with permission from Krause, J., et al., "A complete mtDNA genome of an early modern human from Kostenki, Russia," *Current Biology* 20:231, 2010.

Both the amount of degradation of ancient DNA and the amount of damage are influenced by the environmental circumstances of the specimen. In general, cold and dry conditions favor DNA preservation, while hot and humid conditions do not. The best environment by far for preserving DNA is permafrost or other conditions where the remains are frozen quickly after death and stay frozen. While this is good news for those who work with wooly mammoths and the like, unfortunately most of our ancestors have not had the good sense to die in permafrost or frozen environments, with the exception of Ötzi the Iceman, whose 5000-year-old remains were found melting out of a glacier in the Alps in 1991, and the 4000-year-old Saqqaq individual from Greenland, whose hair yielded DNA that was quite well preserved (Rasmussen et al. 2010). Caves also tend to preserve DNA better than other conditions, probably because cave temperatures tend to stay cool and constant—as long as the cave is not periodically flooded, which unfortunately happens all too often. Still, these are at best general tendencies that influence DNA preservation; there is a lot we do not understand in terms of why particular specimens do or do not yield DNA. Moreover, there can be significant variation in the amount of DNA recovered from different bones from the same individual, or even different parts of the same bone.

It would obviously be very helpful if there was some simple method for predicting whether or not a specimen would be likely to yield DNA. Several years ago, the amount of **aspartic acid racemization** in collagen from skeletal remains showed some promise as a proxy for DNA preservation. All amino acids (the constituents of proteins—see Chapter 2 if you need a refresher) can occur in two forms, called D and L, which are mirror images of each other. It turns out that practically all of the amino acids in living organisms are in the L form, but after death, the L form will spontaneously change to the D form—this process is called racemization. More importantly, racemization of one amino acid, aspartic acid, seemed to occur at about the same rate as DNA degradation (Poinar et al.

1996). Determining the ratio of the D/L forms of aspartic acid is simple, fast, requires just a few grains of bone powder (the amount of bone powder from drilling a small hole is more than ample), and is not so prone to contamination from modern sources (you'd have to handle a specimen pretty roughly to contaminate it with your bone collagen!). It therefore seemed an ideal proxy method for assessing the likelihood of a specimen having sufficient surviving DNA for analysis, and studies made use of aspartic acid racemization to screen specimens for ancient DNA analysis (e.g., Kumar et al. 2000). Alas, further studies of many more samples have discounted the effectiveness of aspartic acid racemization as a proxy for DNA degradation (Collins et al. 2009), and currently there are no effective predictors as to whether or not a particular specimen will yield enough authentic DNA for analysis. You simply have to try and see what you get (which sometimes means convincing a skeptical curator to let you drill a small hole into a precious specimen to recover enough bone powder for DNA extraction and analysis!).

PROPERTIES OF ANCIENT DNA: CONTAMINATION

The final issue that is of concern with ancient DNA is contamination. There are three types of contamination that can potentially influence ancient DNA analysis: (1) substances in an ancient DNA extract that inhibit subsequent analysis; (2) "environmental DNA" (coming from fungi, bacteria, etc.) in the ancient DNA extract; and (3) contamination from modern human DNA introduced by people handling the remains and/or during the laboratory procedures. Regarding the first, substances from the ancient remains that copurify with the DNA can inhibit subsequent manipulations of the DNA extract. This is not a just a problem with ancient DNA; for example, DNA extracts from blood may contain heme (the part of hemoglobin that contains iron and binds oxygen), especially if a lot of the red blood cells in the blood sample have lysed, and heme can inhibit DNA polymerases and thus interfere with PCR or other amplification steps. However, while not limited to ancient DNA, inhibition is a frequent problem with ancient DNA extracts, probably because skeletal and other remains have often been in contact with soils and other environmental elements for up to tens of thousands of years and thus have had the opportunity to soak up all sorts of strange substances. Or, some remains may have been specially treated after the death of the individual (e.g., Egyptian mummies) with preservatives or other substances that have an adverse impact on DNA manipulations.

There are a number of ways to deal with inhibition. The first step is to identify that inhibition really is the problem. Usually, the first indication that there might be inhibition is that you carry out a PCR assay on your DNA extract, but do not get any PCR product. This could be due to inhibition, or it could also be that there simply is too little DNA in the extract for the assay to work (or, it could be that you made a mistake in setting up the PCR assay, which is why the prudent scientist will always include a positive control). The way to distinguish between these two possibilities is to take a little bit of your ancient DNA extract and spike it with a small amount of DNA that you know, from previous experience, will give a good result in your PCR assay. If you now get a PCR product, then it is likely that there was too little DNA (or the DNA is too fragmented) for your PCR assay, but if you still don't get any PCR product, then it is likely that something in the ancient DNA extract is inhibiting the PCR assay.

Once inhibition is identified as the problem, there are a few ways to try to overcome it. Inhibition usually (but not always) reflects some substance in the DNA extract that will bind to any protein and prevent it from having any activity. And since the DNA polymerase used in PCR assays, as well as the ligases/polymerases used in adding adapters for preparing sequencing libraries for next-generation sequencing, are proteins, the inhibitor(s) bind to them. Sometimes, the inhibitor is present in quite small amounts, in which case a very effective and easy remedy against inhibition is to simply dilute the extract before adding it to the PCR assay. It may seem counterintuitive that taking a smaller amount of extract works when a larger amount doesn't, but the idea is that you end up diluting the inhibitor enough that it no longer binds to all of the DNA polymerase, while at the same time you still have enough DNA in the diluted extract for the assay to work. Or, you can try adding another protein to the extract that does not interfere with the assay but can "soak up" the inhibitor by binding it. Bovine serum albumin is commonly added for this purpose, not because it has any special affinities for inhibitors but simply because it is readily and cheaply available and is largely inert in subsequent assays or manipulations of the ancient DNA. When all else fails, you can try repurifying the extract with more stringent conditions to try to get rid of everything else but the DNA—you might think this should be the first thing to try, but there is inevitably a trade-off between the quality and the quantity of DNA in an extract: the higher the quality, the more DNA you lose during the extraction. And with the low DNA amounts that are typical for ancient DNA, it is generally better to use DNA purification methods that maximize yield versus quality, otherwise you risk ending up with a highly purified, high-quality DNA extract that doesn't contain enough DNA for subsequent analysis.

The second form of contamination that can influence ancient DNA is so-called "environmental

contamination," coming from microbes, fungi, and so forth. Again, this is not limited to ancient DNA; while DNA from blood samples should be 100% human DNA, as discussed in Chapter 8, we generally prefer cheek swabs or whole saliva when sampling human populations for genetic diversity studies, and these samples will also contain bacterial DNA from the oral cavity. Still, these freshly obtained samples do contain substantial amounts of human DNA, usually 50% or more. By contrast, ancient DNA extracts can sometimes consist almost entirely of microbial and/or fungal DNA (from organisms that colonized the remains after the death of the individual) and hardly any endogenous DNA (i.e., DNA from the individual who died). For example, the amount of Neandertal DNA in the extracts that were used to determine the Neandertal draft genome sequence (Green et al. 2010) was about 1.5–3%, with the rest coming from bacteria or fungi—and these are the "best" samples that could be found in terms of amount of endogenous DNA!

This low level of endogenous DNA in a sea of microbial/fungal DNA places some limitations on the further analyses that can be done. For PCR assays, there is generally no problem, as long as the primers are specific, but this is the age of genomics, and now we'd like to get more information from our ancient DNA specimen than the few hundred base-pairs or so of sequence that can be obtained via PCR-based approaches. Current practice with modern samples uses shotgun sequencing (i.e., sequencing all of the DNA in an extract) and next-generation sequencing platforms (discussed back in Chapter 7), but this is not cost-efficient when only about 1% of the sequences are what you want—at the current cost of roughly $5000 to obtain a complete genome sequence from a sample with 100% human DNA, it would cost about half a million dollars to obtain the equivalent genome sequence from an extract with 1% endogenous DNA.

With such extracts, one must instead employ procedures to enrich for the desired DNA prior to sequencing. One crude (but effective) way of doing so, that was employed in the Neandertal genome study (Green et al. 2010), is to treat the DNA extract with a restriction enzyme that cuts bacterial/fungal DNA preferentially (see Chapter 7 if you need a refresher on restriction enzymes). Such restriction enzymes recognize sequences that are rich in CG dinucleotides, which are relatively infrequent in human DNA but more common in bacterial DNA (most bacteria have a higher overall GC content than human DNA). Treating the Neandertal DNA extract in this way reduced the fraction of bacterial/fungal sequences from 97% to about 87%—not great, but still an improvement, although one inevitably loses some Neandertal DNA sequences to the restriction enzyme digestion as well.

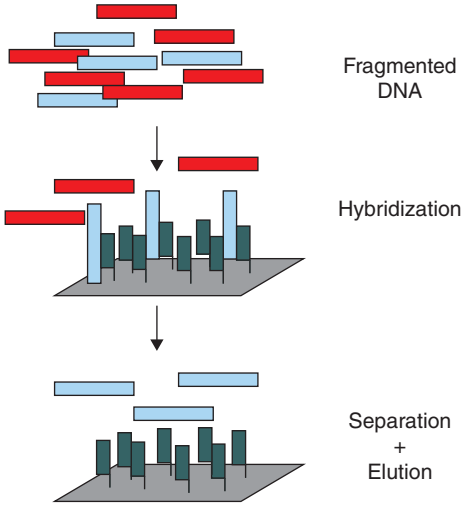


FIGURE 15.4

Capture-enrichment hybridization on an array. Probes to the genomic region(s) of interest (light blue) are immobilized on a solid array. Fragmented DNA containing both the DNA of interest (light blue) and contaminating DNA (red) is hybridized to the array and complementary DNA fragments are bound. The unwanted, unloved DNA (red) is washed away and then the desired DNA is eluted from the array and used to prepare the library for next-generation sequencing. Reprinted with permission from Teer, J.K., and Mullikin, J.C., "Exome sequencing: the sweet spot before whole genomes," *Human Molecular Genetics* 19:R145, 2010.

More recently, hybridization of a DNA extract to an array has proven to be an efficient and cost-effective means of enriching prior to next-generation sequencing (Figure 15.4). These arrays are based on the same general principle as the genotyping SNP chips discussed in Chapter 7, in that the array consists of immobilized DNA probes, and DNA sequences in the extract that are complementary to the probes will bind to the array. An important distinction, though, between this array-based capture and the SNP chips is that whereas the SNP chips are used under conditions where only the specific alleles will bind to the probes, the capture arrays are designed to capture all DNA sequences that are similar to (but not necessarily identical to) the probes, so that sequence variants in the captured DNA can be determined. The noncomplementary, nonbinding DNA is washed away, and then the bound DNA is released from the probes and shotgun-sequenced. Unfortunately, it is not cost-effective (yet) to produce such arrays for the entire human genome, and repetitive regions in the genome must be avoided as well (e.g., if you have probes to an Alu repeat, then you can expect all of the million or so Alu repeats in an individuals' genome to hybridize to the probes, resulting

in a big mess). Still, you can buy arrays for the entire **exome** (i.e., all of the known protein-coding sequences, or exons, in the human genome) or design arrays and have them produced for any genome region of interest—currently one array can enrich for over a million base-pairs, and that will undoubtedly improve in the near future. Issues do arise with using arrays designed from the genome sequence of a modern human to capture DNA from an archaic human such as a Neandertal, as highly divergent sequences may not be captured so well and thus end up underrepresented in the resulting sequences. Still, array-based capture is the current method of choice for enriching for desired target sequences prior to sequencing; however, the technology is changing rapidly, and who knows what the future will bring.

The third type of contamination that can occur with ancient DNA—and by far the most insidious—is contamination with human DNA. This contamination can arise during excavation, retrieval, and cleaning of the remains; subsequent handling in the anthropology laboratory or museum; and/or during the DNA extraction and further manipulation in the genetics laboratory. As just one example of where such contamination can come from, in the time it takes you to read this sentence, you will have shed a few thousand microscopic skin flakes into the air, some of which contain DNA. If these flakes fall onto a bone from which you are extracting DNA, or into a tube that contains your DNA extract, then you now have DNA contamination. And if you were to sneeze or cough on your bone or open tube of extract, well, you can imagine the consequences!

DNA contamination is not such an issue with most contemporary samples that come from blood or saliva, because in each drop of a DNA extract obtained from such "fresh" samples, there will typically be several thousand to several million copies of any given segment of DNA. So, if a few copies of someone else's DNA should get into the extract via skin flakes or whatever, it's not going to make any difference. With ancient DNA, though, where the number of copies of any given segment of DNA in a drop of extract can, in extreme cases, probably be counted on the fingers of one hand, DNA contamination is a much more serious issue. As we shall see later on in this chapter, DNA contamination has led to some rather extravagant claims concerning ancient DNA that later turned out to be quite wrong.

So what can one do about DNA contamination? Unfortunately, only so much. In the laboratory, to be sure, one can take extreme care and precautions. For example, specimens should be handled in dedicated ancient DNA laboratories, in which no modern DNA is allowed to enter (except in the form of the people doing the work, of course!). Everything in the laboratory is treated with ultraviolet light, to destroy any lingering DNA contamination, and the people who work in the laboratory typically wear "space suits" or the equivalent to minimize any DNA coming from

FIGURE 15.5

What it takes to work with ancient DNA. Left, ancient DNA guru Matthias Meyer wearing protective clothing and working in a dedicated ancient DNA laboratory that is kept under positive air pressure (to keep out unwanted contamination) and regularly exposed to ultraviolet light (to degrade any contaminating DNA). Right, if you are really fortunate, you can persuade the archaeologists excavating a site to wear protective clothing. While this doesn't happen so very often—for obvious reasons, try working an archaeological excavation in such clothing!—It does make a big difference in terms of keeping skeletal remains as free of DNA contamination as possible. Left, reprinted with permission from the Max Planck Institute for Evolutionary Anthropology. Right, reprinted with permission from Ko, A.M., et al., "Early Austronesians: Into and out of Taiwan," *American Journal of Human Genetics* 94:426, 2014.

them (Figure 15.5). Extensive controls, which include all chemicals used in the DNA extraction and subsequent assays but no added source of DNA, are routinely included in all of the analyses to make sure that all reagents are free of DNA. In short, DNA contamination arising from the laboratory can be controlled, albeit with a lot of effort and care.

However, a lot of DNA contamination can occur during excavation and subsequent handling of the remains, and for remains that are already in museums, the damage is already done. But for newly excavated remains, there is some hope. Some anthropologists, while excavating potentially important remains, have taken to wearing the "space suit" garb that the laboratory people use in the ancient DNA laboratory. Such precautions may seem excessive but they really do help; in one test carried out by our institute in Leipzig, an anthropologist wearing protective clothing excavated a small bone at a field site, cut it in half, put one half into a sterile bag, took the other half and handled it as would normally be done in the field and put it into a sterile bag, and then shipped both bags to our institute. The bone that had not been handled in the field did not show any signs of DNA contamination, while the bone that had been handled in the field was already contaminated. So, extensive precautions in the field can help.

And what can be done about archival remains that have been extensively handled? At the moment very little—but the situation is not entirely hopeless. Often, such DNA contamination is limited to the surface of the remains, so by removing the surface material and/or drilling into the center of the bone, a relatively uncontaminated (or at least, less contaminated) sample can be obtained—this seems to be especially true for teeth. Moreover, a very recent development that has a lot of people excited is the discovery that the petrous bone (part of the temporal bone, which harbors the inner ear) seems to preserve DNA better than any other part of the skeleton and to be relatively free of contamination (Pinhasi et al. 2015). Otherwise, the best way to deal with DNA contamination is to be able to identify it and distinguish it from the endogenous DNA coming from the remains. Studies of ancient DNA from other creatures, therefore, have a tremendous advantage over the analysis of DNA from humans or our relatives; if you get bison-like DNA out of a bison bone, you can be pretty sure that it did not arise from contamination (unless you are working in a laboratory environment with a lot of bison DNA around—and such things have happened, as we shall see later). But if you get human-like DNA out of a Neandertal bone, is it from the bone, or is it contamination? The more similar the remains are to modern humans, the bigger the problem—as you can well imagine, if you are working with the remains of ancient Europeans, it can be difficult, if not impossible, to distinguish authentic ancient DNA from contamination (especially if the people who have handled the remains and/or work in the laboratory are of European ancestry, which is usually the case). For this reason, many researchers (including myself) were extremely skeptical about many of the results coming from studies of ancient DNA from the remains of recent modern humans—in many cases, they probably reflected contamination.

Fortunately, the new advances in next-generation sequencing promise to facilitate the identification of contamination. As shown in Figure 15.3, there is a characteristic spectrum of C → T and G → A changes, reflecting cytosine deamination, that can be identified in sequence reads that come from ancient, but not modern, DNA. If you sequence DNA from an ancient specimen but you don't see this, then you should be highly suspicious that you're sequencing DNA contamination. Since next-generation sequencing provides many independent reads for each position (similar in principle to the independent clones of PCR products shown in Figure 15.3), heterogeneity in the reads can be another indication of DNA contamination. Thus, with next-generation sequencing, it is proving possible to obtain accurate sequences even from ancient DNA derived from modern human remains, and recently some very large-scale studies have been published (e.g., Allentoft et al. 2015; Mathieson et al. 2015).

One criterion that was proposed in the past for helping to ensure the authenticity of ancient DNA results was independent replication—that is, another laboratory should analyze another sample of the remains and see whether they get the same results or not (Cooper and Poinar 2000). On the one hand, this would seem to be the best way to ensure the accuracy of the results, and after all it is a basic tenet of science that all results should be repeatable by someone else—otherwise, it isn't science. In practice, though, independent replication often raises more problems than it solves. It can be quite difficult to persuade the curator of a valuable fossil that they should let someone else grind up yet more of their precious fossil, not to mention finding another laboratory to carry out a lot of expensive and time-consuming work, just to "confirm" someone else's result. Perhaps the biggest argument against independent replication, though, is that it doesn't always work. Several years ago, a 50,000-year-old bone was analyzed by someone at our institute; they obtained a PCR product to a segment of mtDNA, sequenced clones of the PCR product, and got what seemed to be one human sequence. Someone else repeated the analysis on another sample of the bone and got the same sequence. You would think that an independently replicated human sequence from a 50,000-year-old bone would make headlines and be

a cause for celebration, and ordinarily you would be right—except that the bone in question came from a cave bear! It just goes to show how pervasive DNA contamination can be.

HISTORY OF ANCIENT DNA STUDIES

Ancient DNA as a field got its start in 1984, when Russ Higuchi (working with Allan Wilson at Berkeley) was able to obtain 229 base-pairs of mtDNA sequence from a quagga (Figure 15.6), an extinct relative of horses and zebras (Higuchi et al. 1984). This was a technical tour-de-force, as Russ modified conventional cloning methods, which ordinarily require much more DNA than can be obtained from a 100-year-old piece of skin. And with the resulting mtDNA sequence, Russ was able to resolve an ongoing controversy about the phylogenetic relationships of quaggas, zebras, and horses that was somewhat akin to the chimpanzee–gorilla–human trichotomy question: there were those who thought quaggas were most like zebras, those who thought quaggas were most like horses, and those who thought quaggas were quaggas (i.e., horses and zebras were most closely related, and quaggas were the outgroup). The mtDNA results indicated that the quagga was in fact closely related to the plains zebra; admittedly, while this was not the sort of question that kept people awake at night, it nonetheless nicely illustrates the power of ancient DNA analyses to address questions about the phylogenetic relationships of extinct creatures.

The quagga study was soon followed by a report of a cloned DNA sequence from an Egyptian mummy (Pääbo 1985), which noted paleogeneticist Svante Pääbo obtained when he was a graduate student (and was supposed to be working on a different project). Although it is more likely than not that this putative mummy sequence actually reflects contamination with modern DNA, both the quagga and the mummy study got people seriously thinking about the possibilities of ancient DNA and captured the public's imagination—sometimes, not always for the best. I still recall seeing Russ Higuchi get off the phone in Allan Wilson's laboratory (where I was a graduate student at the time) shortly after the quagga study was published, with a stricken look on his face. When I asked what was the matter, he replied that he had just spoken with someone whose son had died recently, and the father wanted to know whether Russ could use his ancient DNA work to clone his son and bring him back to life. For the record, while there have been great advances in obtaining and sequencing ancient DNA, and some advances in cloning mammals from cells, we are a long, long ways away from being able to recreate a full ancient genome sequence, put it into a cell, and grow a new ancient individual from that cell. As we have seen, ancient DNA contains all sorts of modifications and damage, and sorting out the real sequence from artifacts cannot be done with anything approaching 100% certainty. Moreover, what we call a "complete" genome sequence is currently only about 85% complete, because of limitations with the technology. Not only are there huge technical problems to be surmounted before we could ever even think about cloning a Neandertal (for example), there are also ethical questions to be addressed—even if we could clone a Neandertal, should we? What sort of life would such an individual have? That is something the public (you, for example) will have to decide if and when the day comes that it is technically feasible to clone a Neandertal.

Anyway, while the quagga and mummy studies ignited scientific interest, what really got the ancient DNA field going was the development of PCR. PCR overcame all of the technical limitations of conventional DNA cloning, as with specific primers one could amplify DNA segments of interest from samples with even very minute amounts of endogenous DNA in a sea of bacterial or fungal DNA. These amplified DNA products could then be manipulated in a variety of ways, as discussed in Chapter 7. The late 1980s and early 1990s saw a plethora of ancient DNA studies from remains that were astonishingly old: 17 million-year-old magnolia leaves preserved in shale deposits (Golenberg et al. 1990); 40–120 million-year-old insects preserved in amber (Cano et al. 1993); even 80 million-year-old dinosaur bones (Woodward et al. 1994)—it really seemed as if the sky was the limit when it came to ancient DNA.

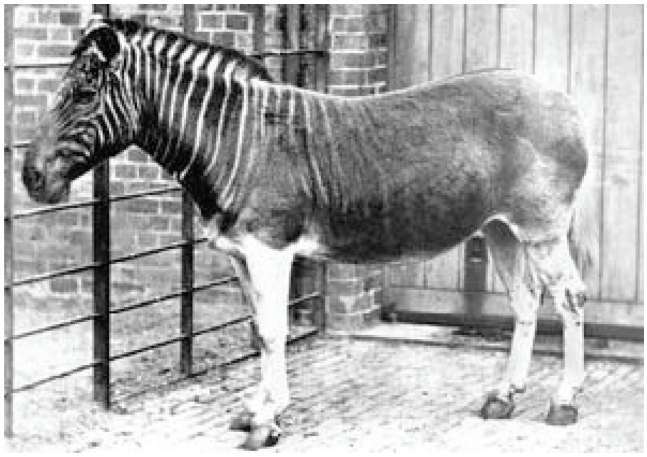


FIGURE 15.6

A quagga photographed in the London Zoo. The quagga was the first extinct creature from which a DNA sequence was obtained. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Quagga_photo.jpg).

Alas, when something seems too good to be true it often is, and so it was with these claims of extraordinarily old ancient DNA. None of the claims of ancient DNA from these remains that are millions of years old have stood up to further scrutiny. The dinosaur DNA turned out to be a weird artifact involving human DNA contamination, while the claims of DNA from magnolia leaves or from insects in amber have not been reproduced in subsequent studies. That is not to say that the original claims were necessarily false—as discussed previously, DNA preservation varies widely, even within the same specimen, so as wildly improbable as it sounds, maybe the studies claiming success were lucky to hit upon the rare specimen with surviving DNA, and nobody else has had the same luck since then. But again, according to the standards of science, if a result is not reproducible, then while it may not be wrong, it is not a scientific result. And the following story is yet another illustration of just how difficult it is to deal with DNA contamination in ancient DNA. When paleogeneticist Hendrik Poinar was a graduate student with Svante Pääbo in the late 1990s, he was given the task of reproducing the ancient DNA results from 40 million-year-old insects in amber. Hendrik was uniquely qualified to do so, as he had participated in some of the original studies of DNA from insects in amber when he was an undergraduate student (Cano et al. 1993), and also because his father, George Poinar, was an expert on amber inclusions and had an extensive collection. Hendrik tried several specimens without any success, until one day he got an insect-like DNA sequence from a mosquito in amber. Alas, the excitement was short-lived, as upon further inspection the DNA sequence turned out not to be related to mosquitos but instead was identical to modern fruit flies—and there was a fruit fly laboratory just down the hall!

And so ancient DNA then went through a period of "doom and gloom," when some experts doubted the veracity of practically all ancient DNA results. But out of this period arose a growing consensus that it was indeed possible to obtain authentic ancient DNA from at least some remains, and informal guidelines were developed and proposed as standards for the field to adopt (e.g., Cooper and Poinar 2000). These guidelines included the practices discussed previously, namely, dedicated ancient DNA laboratories and equipment, extensive negative controls, sequencing of many independent clones of PCR products to investigate contamination, and independent replication as the true "gold standard." Predictably, not everyone agreed with the necessity for these, especially cloning of PCR products and sequencing many clones (which is expensive and time-consuming), and independent replication (which many found problematic for the reasons discussed previously). Still, it is now generally accepted that if one is fortunate enough to have access to remains with preserved DNA, with sufficient care it is possible to generate authentic ancient DNA results. And the next-generation sequencing methods have proved a godsend for ancient DNA, with the signature of cytosine deamination providing a way of authenticating ancient DNA independent of the actual sequence obtained. Thanks to these advances, we now have complete genome sequences from archaic humans such as Neandertals and the like (as discussed in the "Ancient DNA: archaic humans" section). We are also currently in the midst of a revolution in terms of ancient DNA sequences from modern humans, with new studies of up to hundreds of individuals appearing almost every week. Currently, the oldest authentic ancient DNA comes from insects in ice cores from Greenland that are about 450,000 to as much as 800,000 years old (Willerslev et al. 2007) and from a horse bone found preserved in permafrost dating to about 700,000 years ago (Orlando et al. 2013). So, it would seem that a million years or so is the upper bound for obtaining authentic ancient DNA with current methods—but who knows what the future will bring.

ANCIENT DNA: ARCHAIC HUMANS

From the standpoint of anthropology, the most spectacular results to come from ancient DNA studies are the insights from DNA from archaic humans. The first such success was from the Neandertal-type specimen, found by quarry workers in the Neander Valley in Germany ("Tal" is the German word for valley) in 1856. Paleogeneticist Svante Pääbo persuaded the curators of the fossil to let him take a small piece of the upper forearm, and in early 1997, graduate student Matthias Krings set to work on it. I was fortunate to be spending a sabbatical leave at that time with Svante in Munich, and I vividly remember coming a few minutes late to the weekly meeting of the ancient DNA group and finding Svante, Matthias, and the others hunched over a computer printout, clearly very excited. Matthias had extracted DNA from the piece of bone, used PCR to amplify a 60-bp fragment of mtDNA, and then cloned the PCR product and sequenced several clones. The consensus sequence for the clones showed seven nucleotide differences from the human reference mtDNA sequence, which was a lot—most humans would not show any differences from the reference sequence in this 60-bp segment of mtDNA, and the most you would expect would be one or two differences. This immediately suggested that the sequence was not from a modern human but rather was from something clearly related to modern humans. Over the ensuing weeks, Matthias amplified and cloned and sequenced, and amplified

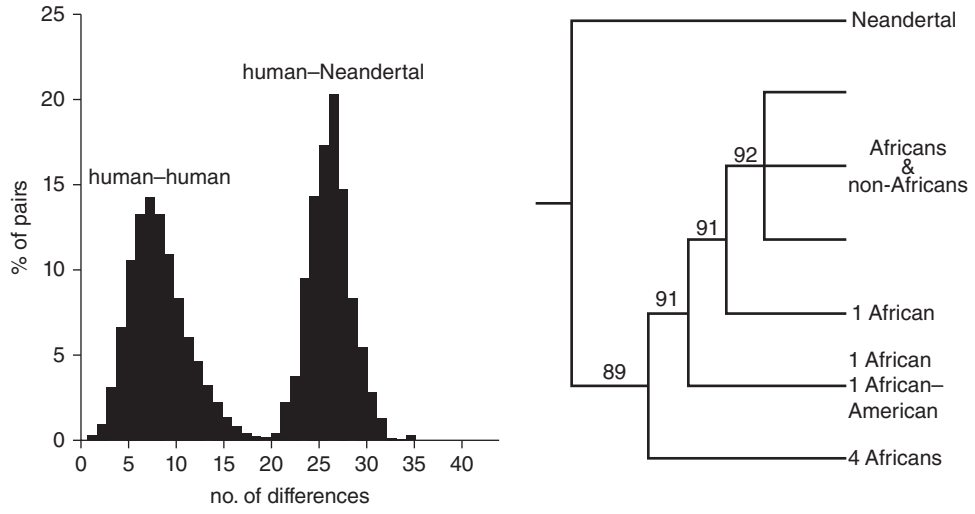


FIGURE 15.7

Comparison of the first mtDNA HV1 sequence from a Neandertal to modern human HV1 sequences. Left, distribution of the number of pairwise differences between two modern human mtDNA sequences or between a modern human and the Neandertal mtDNA sequence. Right, a phylogenetic tree relating the Neandertal HV1 sequence to modern human mtDNA sequences (numbers on branches are bootstrap values). Both analyses indicate that the Neandertal HV1 sequence falls outside the range of modern human HV1 sequence variation. Reprinted with permission from Krings, M., et al., "Neandertal DNA sequences and the origin of modern humans," *Cell* 90:19, 1997.

and cloned and sequenced, and amplified and cloned and sequenced. Many of the individual sequences showed substitutions that looked like cytosine deamination (indicating damage), and some were identical to modern human sequences (indicating some contamination). But eventually Matthias was able to piece together a 367-bp mtDNA sequence that a graduate student in my laboratory back at Penn State University, Anne Stone, was able to replicate a part of from another piece of the bone, thereby satisfying all of the criteria then thought to be important for verifying the authenticity of the results. And this Neandertal mtDNA sequence (Krings et al. 1997) fell well outside the range of modern human mtDNA variation (Figure 15.7); sequences from two modern humans differed on average at ~8 positions, while the Neandertal mtDNA sequence differed on average at ~25 positions from any modern human mtDNA sequence. The conclusion from this result: Neandertals did not contribute their mtDNA to modern humans.

Over the ensuing years, partial mtDNA sequences were determined for other Neandertal remains, and with new advances in sequencing technology, at the time I write this we now have complete mtDNA genome sequences from seven Neandertals (with more in the works), and they all give the same picture as that in Figure 15.7: namely, Neandertal mtDNAs do not fall within the range of modern human mtDNA variation. This might seem to argue against any scenario that invokes a genetic contribution from Neandertals to modern humans in general (and Europeans in particular), such as multiregional evolution or any assimilation hypothesis that includes Neandertals. However, there are two important caveats to keep in mind. First, when analyzing archaic human remains, there is an inherent bias toward calling "authentic" those results that give us something recognizably different from modern human DNA, and calling "contamination" anything that looks like modern human DNA. So, if we had a Neandertal-like fossil that had early modern human mtDNA because of interbreeding between Neandertals and early modern humans, we would most likely dismiss the results as reflecting contamination and not take them seriously. Fortunately, this concern is greatly diminished with next-generation sequencing of genomic DNA, as that has the potential of distinguishing mixed Neandertal-early modern human ancestry from pure Neandertal ancestry with modern human DNA contamination (e.g., by examining DNA damage patterns, as in Figure 15.3). It will be particularly interesting to apply such methods to fossils that have been (controversially) claimed to have mixed ancestry as they supposedly exhibit features characteristic of Neandertals as well as features characteristic of modern humans.

The second caveat is that just because Neandertals did not contribute their mtDNA to us does not mean that they did not contribute any of their DNA to us. It could be that Neandertals and early modern humans interbred, but then Neandertal mtDNA was

subsequently lost over the generations leading to us, either by chance (i.e., genetic drift) or because of selection against it (i.e., perhaps Neandertal mtDNA did not function as well with the modern human nuclear genome). While this seems obvious, and indeed was for the most part generally recognized, nevertheless some tried to make a subtle argument to the effect that the lack of Neandertal mtDNA in modern humans did in fact support the replacement scenario for all of our genome (and honesty compels me to admit that I was one of those making this argument). The argument goes like this: when AMHS first encountered Neandertals, the AMHS presumably had some advantages over Neandertals in terms of tool technology, weapons, degree of social organization, maybe language, and so forth. In such situations involving different groups of modern humans, the usual outcome is that males from the more technologically advanced group have access to females from the less technologically advanced group but not vice versa. This is known as **hypergyny** and has been well documented among human groups, for example, when Bantu farmers encountered Pygmy hunter-gatherers in Africa, the Bantu males took Pygmy females as wives but not vice versa; or when Europeans first came to the New World, practically all of the interbreeding between Europeans and native Americans involved European males and native American females. So, by the same logic, when AMHS encountered Neandertals, presumably any interbreeding would have involved AMHS males and Neandertal females, and such matings should, therefore, enhance the contribution of Neandertal mtDNA to AMHS—that is, if Neandertals contributed any DNA to AMHS, it should have been their mtDNA. The subtle inference, therefore, is that since we don't see any such Neandertal mtDNA in AMHS, there was no interbreeding between Neandertals and our ancestors.

Anyway, regardless of what you think of this argument (and I have to confess that I found it rather attractive!), clearly the real way to test it is to somehow obtain the genome sequence from a Neandertal and then see if there is any evidence for Neandertal DNA in AMHS. And for many years that proved to be a real stumbling block. Analyses of ancient DNA in general, and Neandertal DNA in particular, were largely restricted to mtDNA because of the higher copy number of mtDNA—on average, hundreds to thousands of copies per cell versus just two copies of any autosomal DNA segment. Thus, if any ancient DNA survived, the odds favored mtDNA over autosomal DNA. But when you then consider that the average PCR-based mtDNA assay consumes on average 1/10 to 1/50 of the total DNA extract, which typically comes from 0.25–0.5 g of bone, and that there are (at most) a few mtDNA copies in the amount of DNA extract added to the PCR assay, then—well, you can do the math, you'd have to grind up an entire fossil to get enough DNA for just one PCR-based autosomal DNA assay, and it would be rather difficult to convince a curator to let you do that.

To be sure, there have been some successes with PCR-based assays of autosomal loci in Neandertals. For example, PCR-based assays have shown that Neandertals carried mutations that were likely to lead to red hair (but different mutations than those associated with red hair in modern humans), as well as unexpected results concerning the *FOXP2* gene (associated with language capabilities) in Neandertals—more on this in Chapter 17. But the next-generation sequencing platforms are what really made it possible to even think about an archaic human genome sequence, because they are tailor-made for ancient DNA. Next-generation sequence read lengths are short, usually 36–100 bases, which is the size range of most ancient DNA; in fact, fresh DNA usually has to be sheared down to the optimal length for next-generation sequencing (cf. Figure 9.3), but with ancient DNA you can skip that step. Next-generation sequencing involves sequencing each position multiple times, which then naturally provides the information on potential contamination or substitutions indicative of damage that one normally gets only by the tedious process of cloning PCR products and sequencing lots of clones. Svante Pääbo immediately realized the possibilities offered by the first next-generation sequencing platforms and audaciously proposed in 2006 to sequence the complete genome of a Neandertal. Many technical difficulties had to be overcome, and there were times when it seemed like it wouldn't work after all, but in the end everything came together, with the publication in 2010 of a draft Neandertal genome (Green et al. 2010) with an average coverage of 1.3X (remember, this means that each base was sequenced on average 1.3 times—in actuality, about 65% of the genome was sequenced at least once, so 35% was not sequenced at all).

And so what does this Neandertal draft genome sequence tell us about the genetic contribution of Neandertals to AMHS? Three different analyses were carried out to address this question, but since all three gave the same result, we will focus on just one of these, because it is the easiest to explain and understand. This is the D statistic, described back in Chapter 12 (specifically, Figure 12.22). To refresh your memory, recall that with the D statistic, we compare two human sequences to (in this case) the Neandertal sequence and look for polymorphic sites where one human has the ancestral allele (as defined by comparison to chimpanzee and/or other outgroup sequences) while both the other human and the Neandertal have the derived allele. A significant excess of sharing of derived alleles between the Neandertal sequence and one of the human sequences (relative to the other) is an indication of gene flow from Neandertals to the ancestors of that human. However, if both humans always share the same number of derived alleles with the

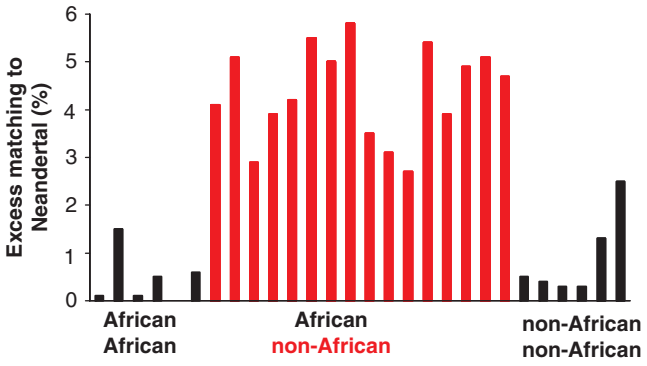


FIGURE 15.8

Excess matching of non-African genome sequences to the Neandertal genome sequence, relative to African genome sequences. The figure shows comparisons between two non-African sequences, between two African sequences, and between an African and a non-African sequence; comparisons in red are significantly different from zero. Data for this figure are from Green, R.E., et al., "A draft sequence of the Neandertal genome," *Science* 328:710, 2010.

Neandertal, then this would indicate no differential contribution of Neandertals to modern humans (i.e., either no Neandertal ancestry or the same amount of Neandertal ancestry in all modern humans, but the latter does not seem very likely given that Neandertals were found exclusively in Eurasia, as far east as the Altai region of Central Asia).

The results obtained via the D-statistic are shown in Figure 15.8: there is no significant excess of sharing of derived alleles by one human versus another when two Africans are compared, or when two non-Africans are compared, but when an African is compared to a non-African, there is always a significant excess of sharing of derived alleles between the Neandertal and the non-African. Surprisingly, this result holds regardless of where the non-African comes from; Papua New Guineans and Native Americans, for example, show just as much Neandertal ancestry as do Europeans, even though Neandertals never inhabited New Guinea or the New World. The implications of this result for human migrations will be explored in the next chapter; for now, the take-home message (corroborated by all three analyses that were carried out to look for Neandertal DNA in modern humans) is that if you are a non-African, then about 2% of your DNA comes from Neandertals.

However, the DNA that we non-Africans share with Neandertals did not necessarily have to come from our ancestors interbreeding with Neandertals—there is an alternative explanation involving ancient population substructure, illustrated in Figure 15.9. The idea is as follows: suppose there was ancient population substructure in Africa, meaning very limited gene flow and hence large genetic differences among populations. Around 600,00–800,000 years ago or so, some archaic humans left Africa, became genetically isolated, and evolved into Neandertals. Then, several hundred thousand years later, the direct descendants of the African population that gave rise to Neandertals gave rise to the AMHS who migrated out of Africa. Later, gene flow across Africa erased the old substructure, and the end result is that non-Africans share a genetic signal with Neandertals because they are both derived from the same ancestral African population. If this seems like a rather clunky scenario, well, it is, but the important point is that the D-statistics cannot

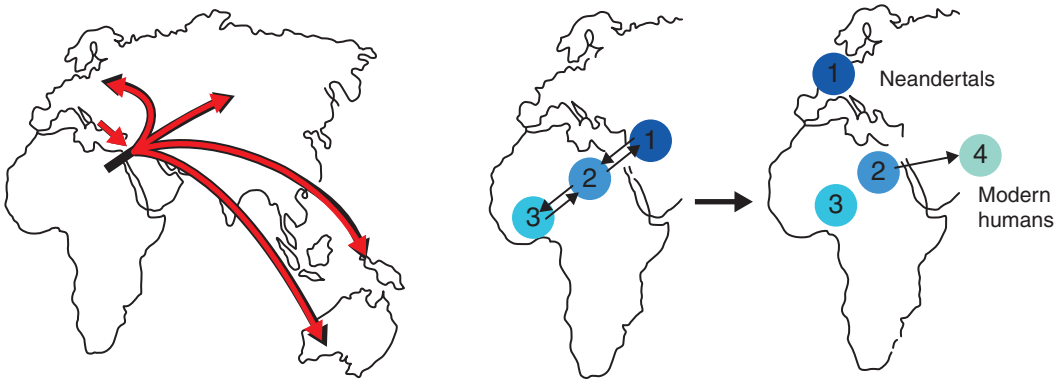


FIGURE 15.9

Two explanations for the apparent signal of Neandertal DNA in all non-African modern humans. Left, admixture from Neandertals into a population that was ancestral to all non-Africans. Right, ancient substructure within Africa. If population 2 gave rise to Neandertals (population 1), then became somewhat isolated from other African populations (population 3), and then gave rise to modern humans (population 4), then this could also explain the apparent admixture signal between Neandertals and modern humans. Reprinted with permission from Sankararaman, S., et al., "The date of interbreeding between Neandertals and modern humans," *PLoS Genetics* 8: e1002947, 2012.

distinguish between ancestral population structure and interbreeding with Neandertals as the explanation for the signal of Neandertal DNA in non-Africans. However, other statistics can. In particular, note that these two explanations make different predictions for the age of the Neandertal DNA in non-Africans: ancient substructure would imply that the signal is hundreds of thousands of years old (corresponding to when the ancestors of Neandertals left Africa), whereas interbreeding with Neandertals implies that the signal is more like 50,000 years old (corresponding to when AMHS left Africa). And very recently, David Reich and colleagues attempted to date the age of the Neandertal DNA signal in non-Africans, based on patterns of linkage disequilibrium, and concluded that it was on the order of 40,000–80,0000 years ago, not half a million years ago. Moreover, the finding of a signal of DNA from another group of archaic humans in AMHS, which we will now turn to, renders the scenario of ancient population structure even more unlikely.

This second signal of DNA from archaic humans in AMHS is an example of pure serendipity. For the Neandertal Genome project, Svante Pääbo's group was screening every potential Neandertal fossil they could get their (suitably sterilized and gloved!) hands on to try to find Neandertal fossils that would have both high amounts of endogenous DNA and low amounts of contamination with modern human DNA. During the course of this work, they determined the mtDNA sequence from a pinky (fingertip) bone from Denisova Cave in southern Siberia. Fingertip bones do not have any identifying characteristics, so it could have been from a Neandertal or from an early modern human, and using next-generation sequencing was the quickest way to see how much Neandertal versus modern human DNA there was in the specimen. To their astonishment, the answer to the question as to whether the pinky bone was from a Neandertal or from a modern human was "neither": the mtDNA genome sequence fell outside the range of both modern human and Neandertal mtDNA variation (Krause et al. 2010), with an estimated divergence date of about 1 million years (Figure 15.10). Determining the complete genome sequence of the pinky bone immediately

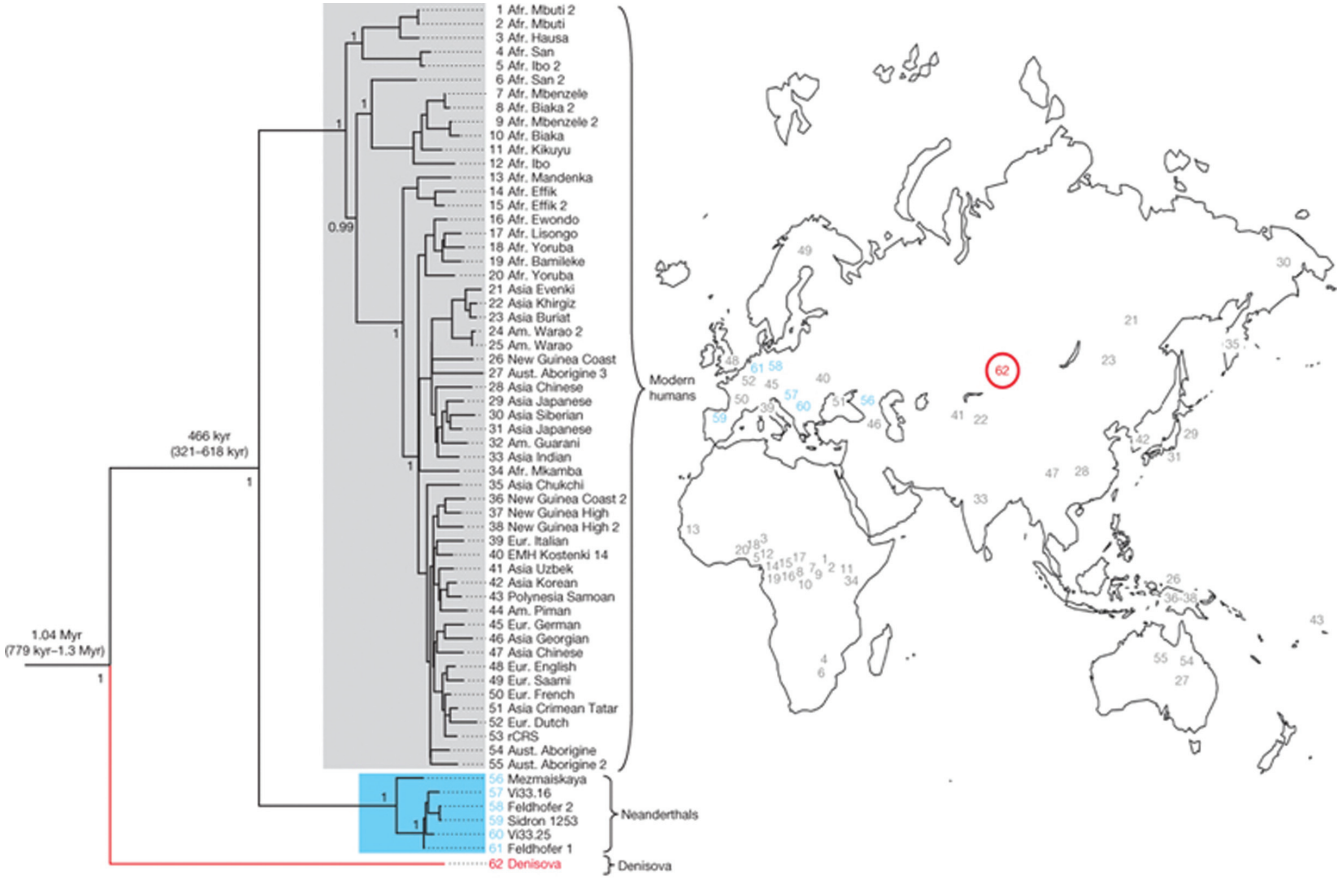


FIGURE 15.10

Phylogenetic tree of complete mtDNA genome sequences from Denisova, Neandertals, and modern humans, along with a map showing where the samples are from. Numbers on branches are bootstrap values. Reprinted with permission from Krause, J., et al., "The complete mitochondrial DNA genome of an unknown hominin from southern Siberia," *Nature* 464:894, 2010.

became a high priority, and as a testament to how rapidly the technology had progressed, a 1.9X draft genome sequence was produced, analyzed, and published (Reich et al. 2010) in the same year as the Neandertal genome sequence. To be sure, one aspect that made the production of the sequence much easier than for the Neandertal remains was that the amount of endogenous DNA in the pinky bone was an amazing 60% (compared to the 3% or so for the best Neandertal remains)—why the pinky bone should have such extraordinary DNA preservation, nobody knows, and it certainly is not the case for other remains recovered from Denisova Cave. But such a high level of endogenous DNA made it feasible to carry out shotgun sequencing without any need for enrichment. Plus, the level of modern human DNA contamination was quite low, probably because anthropologists don't find pinky bones very interesting or informative, and so the bone had not been handled very much.

While the mtDNA results indicated that the pinky bone mtDNA sequence fell outside the range of both Neandertal and AMHS mtDNA variation (Figure 15.10), the genome sequence told a different story (Figure 15.11): namely, the pinky bone genome sequence grouped with the Neandertal genome sequence. This discrepancy nicely illustrates why one should be cautious in drawing too sweeping conclusions based on just a single locus such as mtDNA, as discussed back in Chapter 9. There ensued some discussion as to what to call the individual from

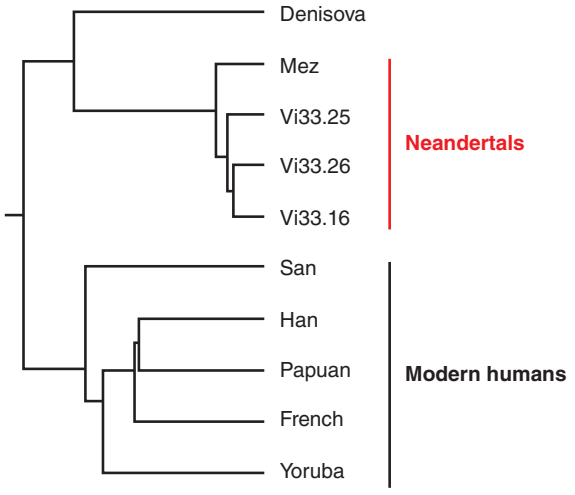


FIGURE 15.11

Phylogenetic relationships of Denisova, Neandertals, and modern humans, based on genome sequences. In contrast to the mtDNA sequences (Figure 15.10), the genome sequences indicate that Denisova is a sister group to Neandertals. Reprinted with permission from Reich, D., et al., "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature* 468:1053, 2010.

which the pinky bone genome sequence was derived, with some advocating it should be a different species, while others felt it was "just" another Neandertal. The consensus, though, was not to make any formal species declaration one way or another—after all, anthropologists have had the remains of Neandertals for more than 100 years, with several hundred remains from lots of individuals, and there is still no consensus as to whether or not Neandertals should be designated as a separate species from us. Instead, the decision was made to simply call them Denisovans (just as Neandertals are named after the place they were first discovered) in recognition of the fact that while they may be related to Neandertals, they do not have exactly the same history as Neandertals. Note that based on the genome sequence, the relationship between Neandertals and Denisovans is slightly more distant than the deepest divergence among modern humans (Figure 15.11).

Currently, all we know about Denisovans is from the genome sequence from the pinky bone, plus two molar teeth from the same cave. These molar teeth are quite extraordinary, as they are huge in size and carry features not seen in the molars of either early modern humans or Neandertals—in fact, in some respects they look more like australopithecine teeth! Still, their mtDNA and partial nuclear genome sequences have been determined (Sawyer et al. 2015), and they do group with the pinky bone for both mtDNA and the nuclear genome, so by that criteria they are Denisovan teeth. Moreover, these differences in tooth morphology also argue that Denisovans deserve their own designation and should not be classified as just a different kind of Neandertal.

The question then naturally arises, is there any evidence that modern humans also interbred with Denisovans? And the really surprising answer that came from the analysis of D statistics was that yes, there is a signal of Denisovan ancestry in some human populations—but only in Melanesians from New Guinea and Bougainville (Reich et al. 2010)! Given that these islands are some 7000–8000 km from Denisova Cave, this is not exactly where anyone expected to find evidence of Denisovan DNA in modern human populations—you could have won a lot of money by betting on this result. The overall implication from the Neandertal and Denisova genome sequences thus is that our ancestors interbred with different archaic humans: all non-Africans carry a signal of Neandertal ancestry, and Melanesians carry an additional signal of Denisovan ancestry. Moreover, this information is of more than just prurient interest for those who want to know about the sex lives of our ancestors; as we shall see in subsequent chapters, the signals of Neandertal and Denisovan DNA in our genomes provide insights into the migration history of

AMHS, as well as into the genetic changes that distinguish us from archaic humans and the genetic adaptations that allowed us to colonize more of this planet than any other species.

Genome sequences from archaic humans are a thriving business, thanks to numerous recent technical innovations in the retrieval and analysis of DNA from fossils; as I write this, we have a high-quality genome sequence from a Neandertal toe bone from Denisova Cave (Prüfer et al. 2014), along with a similar high-quality genome sequence from the Denisovan pinky bone (Meyer et al. 2012), as well as partial genome sequences from two additional Denisovan molars from Denisova Cave (Sawyer et al. 2015). On the horizon—probably before this book is published—are at least one additional high-quality Neandertal genome sequence, and several partial Neandertal genome sequences. Moreover, a recent study has documented extensive Neandertal ancestry in a modern human fossil from Romania called Oase 2; this individual probably had a Neandertal great grandparent, but the Neandertal ancestry in Oase 2 is not related to that in modern humans (Fu et al. 2015). Thus, there were multiple events of Neandertal admixture with modern humans (no surprise there!), not all of which can be detected in modern humans today. The current version of admixture events between modern and archaic humans is now thought to be much more complex than the above simple picture of one admixture event between Neandertals and the ancestors of non-Africans and one admixture event between Denisovans and the ancestors of Melanesians (Figure 15.12); what the current version actually is, and what this tells us about the migration history of our species, will be discussed in the next chapter.

As if all of these archaic genome sequences weren't exciting enough, one of the biggest surprises came from the mtDNA sequence that was recently obtained from what is currently the oldest hominin fossil to yield authentic DNA, namely, a 400,000–500,000-year-old femur bone from the site of Sima de los Huesos in Spain. Morphological analyses suggest that the remains at Sima are related to Neandertals, but the mtDNA genome sequence (Meyer et al. 2014) is instead related to the Denisovan mtDNA sequence (which, you will recall from Figure 15.10, is a really strange, divergent sequence)! To make things even more interesting, in a technical tour de force, ancient DNA guru Matthias Meyer and colleagues have recently been able to obtain a few million bases of nuclear DNA sequence from this fossil (Meyer et al. 2016), and it does indeed group with Neandertals. How to explain the distribution of this very odd divergent mtDNA genome sequence, from Spain to southern Russia, and in different groups of archaic humans (Neandertals and Denisovans), remains a challenge.

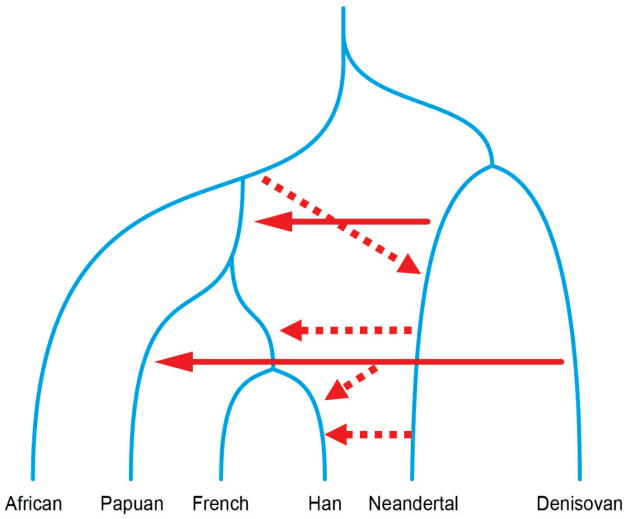


FIGURE 15.12

Current state of knowledge (as of April 2016) concerning episodes of introgression from archaic humans into modern humans (and vice versa). The tree shows schematically the relationships of Neandertals, Denisovans, Africans, Papuans, French, and Han Chinese. The solid red arrows indicate what was inferred right after the determination of the Neandertal and Denisovan genome sequences, namely, genetic contributions from Neandertals to the ancestors of all non-Africans and from Denisovans to the ancestors of Papuans. The dashed red arrows indicate additional inferred episodes of introgression between archaic and modern humans.

The one thing that is certain is that the more we learn about our evolutionary history, the messier it gets.

There are two more points to be made concerning the interbreeding between archaic and modern humans. First, we have genome sequences from two groups of archaic humans—one of which we did not even know existed until we had the genome sequence—and both interbred with early modern humans. The obvious question that this raises: how many other archaic humans were out there that our ancestors interbred with? Indeed, recent studies have claimed that there are signals of some other archaic admixture in the genomes of Africans (Hammer et al. 2011; Hsieh et al. 2016), but this is based exclusively (so far) on DNA analyses of modern Africans, and so other explanations for these signals (selection, outliers due to genetic drift, etc.), while unlikely, are not completely excluded—one really needs the archaic human genome sequence to be sure. Similarly, some segments of the Denisova genome are interpreted as evidence for interbreeding between Denisovans and some as yet unknown, superarchaic "ghost" population (Prüfer et al. 2014). It does seem quite likely that Neandertals and Denisovans were not the only archaic humans

that early modern humans encountered and interbred with, so it will be exciting to see what the future holds in terms of additional discoveries.

Second, there are those who find nothing surprising in these signals of archaic human DNA in modern humans. After all, if you look at the mating habits of modern humans in general—and young human males in particular—it would be surprising if our ancestors did *not* interbreed with any archaic humans they came across. But there is more to this story than randy young males spreading their genes; keep in mind that in order for archaic human DNA to show up in our genomes today, the offspring of matings between archaic humans and early modern humans must have been raised in the early modern human community. This in turn implies that archaic females (or, perhaps, archaic males, although this seems less likely) were accepted into the early modern human communities, were seen as acceptable mates, had children with modern humans, and these "hybrid" children were also accepted into the early modern human communities. Of course, one could postulate other scenarios, such as modern human males raiding archaic groups and forcibly kidnapping the females—there is ample historical evidence of modern human groups engaging in such raids against other modern human groups—but again, the children of such matings between modern and archaic humans must have then been incorporated into the modern human groups. One implication of the archaic human DNA in our genomes, then, is that there was some degree of social interaction between archaic and early modern humans, more so than implied by the scenarios of warfare and conquest that some have envisioned for the spread and ultimate success of modern humans and the concomitant demise of archaic humans. Moreover, the offspring of matings between modern and archaic humans must have been at least partially, if not fully, fertile, in order for the archaic human genes to have been incorporated into the modern human gene pool (i.e., us). This, in turn, implies that archaic and modern humans were all part of the same species, at least according to the biological species concept, which holds that if two individuals are able to have viable and fertile offspring, then they belong to the same species. So, the overall message from the signals of archaic human ancestry in modern human genomes is that the similarities in the behavior and biology of modern and archaic humans must have outweighed any differences.

OTHER USES FOR ANCIENT DNA

While the genome sequences of archaic humans are undoubtedly the most spectacular contribution of ancient DNA to molecular anthropology, they are by no means the only application of ancient DNA. During the "golden era" of ancient DNA (i.e., the late 1980s to early 1990s), when it seemed that DNA could be retrieved from any old remain, it was envisioned that ancient DNA would revolutionize anthropology by allowing genetic comparisons of ancient communities to current communities. As we have already seen, much of molecular anthropology relies on reconstructing the genetic history of populations based on current samples, which thereby has the implicit assumption that the contemporary populations are directly descended from the past populations from that same part of the world. Given the propensity that modern humans have for moving around (and spreading their genes), this is a dubious assumption at best, but one that we can't really do anything about—unless we have ancient remains from the same area, from which we can get DNA. If that is the case, then we can directly compare the DNA variation in the ancient and contemporary populations and see just how representative the latter is of the former.

Alas, while the idea sounds great, the execution initially suffered from the twin evils of too few samples with sufficient surviving DNA to be very informative about the genetic variation of the ancient population, plus the specter of contamination making any ancient DNA results from modern humans highly suspect. All that has changed, and just in the past 2 years or so. Technical advances in the extraction of DNA have greatly increased the chances of success of getting DNA from human remains (especially those that are less than 10,000 years old or so and come from relatively cool climes such as Europe or the Arctic), and even better, with next-generation sequencing ancient DNA comes with a built-in signal of authenticity in the form of excess CG-TA changes at the ends of sequences due to deamination of cytosine (cf. Figure 15.3). If you see this signal in your sequences, you can be pretty sure that you are dealing with authentic ancient DNA; conversely, if you don't see this signal, chances are pretty good that you are sequencing contamination. And if you should find both, indicating the presence of both authentic ancient and contaminating modern DNA, you can fish out the sequences with evidence of cytosine deamination at the ends, as these would be the authentic ancient DNA, and restrict your analyses to these sequences. These technical advances have led to a huge increase in both the number of publications of genome sequences from ancient remains and the number of samples per study—it is not unusual to see studies that analyze dozens or even hundreds of ancient samples (e.g., Allentoft et al. 2015; Mathieson et al. 2015), which would have been unheard of just a few years ago. Ancient DNA has come into it's own as an important source of exciting new insights into human population history,

and we'll see further examples of this in subsequent chapters.

Ancient DNA methods have also assisted in forensic cases, especially those involving the identification of remains where the DNA is in particularly bad shape because of the age and/or subsequent treatment of the remains. Some of the most notable successes have involved the identification of remains of historical interest, such as the Romanov Tsar Nicholas II and his family, killed in Ekaterinburg in 1918 by the ruling Bolsheviks to prevent any possibility of their rescue by the approaching White Army. Their bodies were partially burned, treated with acid, and buried. The bodies were discovered in the late 1970s by amateur archaeologists but kept secret until the political environment improved with the breakup of the Soviet Union, and the remains were officially "found" in 1991. DNA testing was then carried out, and comparisons to living relatives of the Romanovs (including Prince Philip, the Duke of Edinburgh) confirmed the identity of the remains (Gill et al. 1994)—although the remains of two of the five children were missing, fueling speculation that they had survived the killings. I was later involved in DNA testing of Anna Anderson, who had claimed to be Anastasia, one of the Romanov daughters. During her lifetime, Anna Anderson was interviewed by several people who had known the Romanovs, some of whom came away convinced that she was indeed Princess Anastasia, while others came away equally convinced that she was a fraud. Anna Anderson died in 1984 with her claim unresolved; DNA testing was later carried out on a tissue biopsy specimen (from a hospital archive) as well as on hairs (tucked away in an envelope found in a book that had been purchased from her estate). Alas, her fairy-tale story, which captured public attention and was portrayed in Hollywood movies and the like, proved to be an invention of her imagination, as the DNA testing confirmed that she was not a Romanov but instead was a Polish factory worker with a history of mental illness who had disappeared in a munitions factory explosion in Berlin in 1918 (Gill et al. 1995). And subsequent discoveries of the remains of the two missing Romanov children were verified via DNA testing (Coble et al. 2009), putting to rest other such claims. In sum, ancient DNA methods have proven useful whenever DNA is highly degraded, present in very limited amounts, or both, although it is still an open question as to how old remains have to be in order for the term "ancient DNA" to apply—I, for one, am not so amused when I read about "ancient DNA" from remains that are younger than I am!

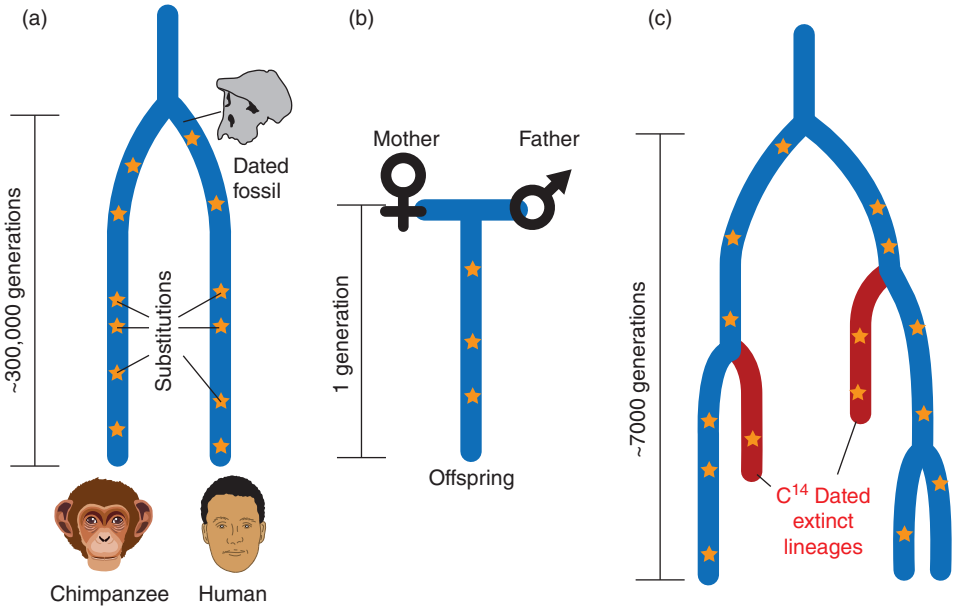


FIGURE 15.13

Three ways to calibrate the rate of molecular evolution. (a) The traditional approach, comparing the number of substitutions between two species with a securely dated fossil record. (b) The direct approach, counting the number of new mutations in a child. (c) A novel approach based on the number of "missing" mutations in DNA sequences from securely dated fossils. In the example shown, the fossil sequences have two fewer mutations than contemporary sequences, and this information can be used (along with the date of the fossils) to estimate the rate of substitution over time. Reprinted with permission from Green, R.E., and Shapiro, B., "Human evolution: turning back the clock," *Current Biology* 23:R286, 2013.

In addition to questions about the continuity of populations over time, ancient DNA can also be used to address questions about the ages of particular mutations. We have already seen in Chapter 12 how we can date the age of a mutation from the amount of linked variation, but keep in mind that all such methods invoke various assumptions, some of which may be quite wrong (such as a constant population size over time). Ancient DNA can, in the right circumstances (i.e., having fossils of the right age with enough DNA for analysis), provide a convenient reality check on such age estimates. If we estimate that a particular mutation occurred 50,000 years ago, then obviously we'd better not find it in a fossil that is 100,000 years old, or something is wrong. In fact, in Chapter 17, we will see an example in which ancient DNA contradicted an age estimate based on variation in contemporary populations, involving important mutations in a gene called *FOXP2* that are associated with human speech abilities.

There are many other possible applications of ancient DNA to molecular anthropology, with more coming up all the time, too many to mention in detail. As just one very recent example, Figure 15.13 shows how mtDNA sequences from directly dated fossils can be used as a novel calibration of the rate of mtDNA evolution. In addition to studies of humans (and our relatives/ancestors), there is a lot to be learned from studies of other creatures—and such studies have the desirable property that contamination with human DNA can be readily distinguished from the authentic ancient DNA. For example, the domestication of animals was a key event in human evolution, and ancient DNA has proven useful in tracing the origin(s) and spread of various domesticated beasts (e.g., Larson et al. 2012; Ottoni et al. 2013; Schubert et al. 2014), as well as promising to provide insights into the genetic changes that were important during the domestication process (Flink et al. 2014). Studies of ancient parasites and disease-causing organisms are also proving useful in further understanding the origins and crucial role of particular diseases in shaping our past population dynamics (Bos et al. 2011, 2014). In sum, thanks to recent and ongoing improvements and developments in DNA extraction and sequencing technologies, the outlook for ancient DNA has never been better, and in the next few years we can expect lots more to come from ancient DNA—although we should still not expect to ever see DNA retrieved from dinosaur bones or from insects in amber!

CHAPTER 16

DISPERSAL AND MIGRATION

Having established that our species arose in Africa, we can now ask, so then what happened? There are two major types of processes, that in some sense define what it is to be a modern human, that concern us. The first is the dispersal and migration of people both across and out of Africa, initially across the Old World, but ultimately to the farthest corners of the globe. No other species—with the exception of our parasites—has reached all the places we have. The second process consists of the genetic adaptations that accompanied the origin and spread of modern humans. In this chapter, we will consider human dispersals and migrations, while in the next two chapters, we will discuss the role of selection and adaptation.

Incidentally, for those who are curious about the distinction between "dispersal" and "migration," in biology the former usually refers to small-scale movements of animals within a particular habitat, while the latter refers to longer-distance directional movements from one habitat to another, often associated with seasonal changes. For example, geese spend the summer months in northern climates, dispersing among various lakes, meadows, and so forth, in a particular area, but then in the fall they undertake a long-range migration to warmer climates in the south, where they spend the winter. For prehistoric humans, this distinction doesn't make much sense—it is generally impossible to know whether people were moving only over short distances each generation and gradually expanding their range over time, or whether they deliberately set out to move long distances in a short period of time—so we will use the terms "dispersal" and "migration" interchangeably.

A thorough description of all of the various migrations that have occurred during the course of human prehistory is beyond the scope of this book—it would entail a book in its own right (and indeed already has, see *The History and Geography of Human Genes*, or the popular version *Genes, Peoples, and Languages*, both by Luca Cavalli-Sforza and colleagues, for a detailed description of human migrations as told by studies of classical genetic markers). We will therefore focus on three examples of human migration in this chapter, and what genetic analyses have revealed about them. These are: (1) the initial out-of-Africa dispersal of modern humans; (2) the colonization of the New World; and (3) the colonization of the Pacific. In doing so, we will also make use of archaeological and linguistic evidence, as these provide a useful source of comparative evidence on human migrations. In fact, because both archaeology and linguistics have been around longer than genetics, it is usually the case that various hypotheses concerning a particular migration already exist based on archaeological and/or linguistic evidence, and so it is then useful to see what genetics can add to the picture (as we shall see, quite a lot!). So, before we get into the examples, let's first briefly go through a few points to keep in mind about archaeological and linguistic evidence.

The key contribution of archaeology to studying human history and migrations is that it provides dates for the presence of humans (either directly, through their remains, or indirectly, through tools, pottery, or other cultural items) that at least in theory are more accurate than any dates obtained via the molecular clock approach. This is because dating of archaeological items is based on radioactive decay of one isotope into another, which occurs at an absolute, fixed rate, whereas molecular clock dating, as we have seen, is based on the accumulation of new mutations, which is an inherently random (stochastic) process. For example, the most relevant type of dating for our purposes is **C14** (carbon-14) dating, which is based on the radioactive decay of an isotope of carbon called ^{14}C. The ^{14}C isotope occurs naturally in the atmosphere as a (more or less) fixed proportion of the total amount of carbon; atmospheric carbon combines with oxygen to form carbon dioxide, which is taken up by plants via photosynthesis, and then by other living things (including us) by eating plants or by eating creatures which have

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

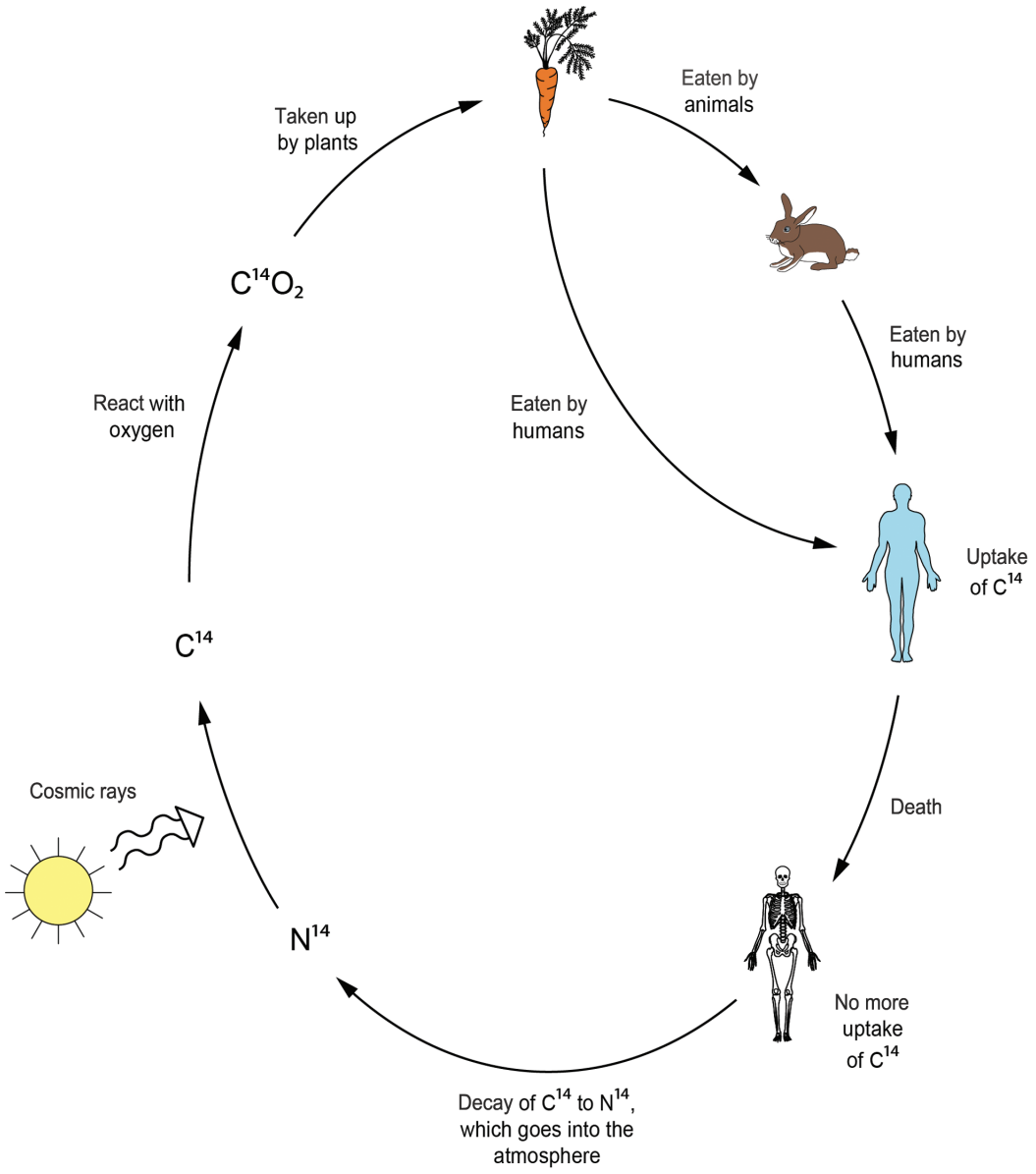


FIGURE 16.1

Outline of the flow of carbon from the atmosphere to living and then dead things and then back to the atmosphere. Starting at the bottom left, nitrogen (N^{14}) in the atmosphere is converted to an isotope of carbon (C^{14}) by cosmic rays, which then reacts with oxygen to form carbon dioxide. This is taken up by plants, which are then either directly eaten by humans or eaten by other animals that are then eaten by humans, with the result that we end up with C^{14} in our bodies. After death, we no longer take in C^{14}, so the C^{14} undergoes decay at a constant rate to N^{14}, which then goes into the atmosphere and so begins the cycle anew.

eaten plants (Figure 16.1). So, while you are alive you maintain a constant proportion of ^{14}C to the usual isotopic form of carbon, namely ^{12}C, because you are constantly taking in more carbon (which is one definition of what it means to be alive), but once you die, the carbon in your body is no longer replenished, and the ^{14}C decays irreversibly to an isotope of nitrogen. The rate at which this happens is expressed as the **half-life**, meaning the amount of time it takes for half of the isotope to decay. ^{14}C has a half-life of 5730 years, so after 5730 years half of the initial amount of ^{14}C will remain, after another 5730 years, a quarter of the initial amount will remain, and so forth. For those of you who like to keep track of such things, this rate of decay amounts to about 14 disintegrations per gram of carbon every minute. Comparing the ratio of ^{14}C to ^{12}C in organic remains (bone, charcoal, shells, etc.) from an archaeological site to the ratio expected in a living

organism thus provides you with the date associated with the death of the remains.

Radiocarbon dating was the brainchild of the chemist Willard Libby, who won a Nobel Prize for his efforts—which were quite considerable. Among other tasks, Libby had to demonstrate that there was indeed detectable radioactivity in carbon derived from living matter, which was accomplished by experimenting on methane gas derived from sewage. And, in order to demonstrate that C14 dating gave reliable results, Libby arranged to test Egyptian mummies from the University of Chicago, for which there were good historical dates. However, one of the first C14 dates he obtained was only a few years ago, instead of the expected many centuries ago, which almost caused him to give up the whole enterprise as one of those good ideas in principle that just don't work out. Fortunately, he decided to test a few other samples, all of which gave the expected dates—and it turns out that the aberrant specimen was a modern forgery!

Because C14 (and other radiometric) dating relies on a fixed decay rate, it provides dates that are quite precise, usually with error ranges of at most a few hundred years. For this reason, some archaeologists make rude comments about molecular clock dating, for which error ranges are typically measured in thousands of years. However, there are some issues concerning C14 dating—and archaeological dating in general—that you should be aware of. The practical upper limit for C14 dating is about 50,000 years—beyond that, there is too little ^{14}C to measure reliably. Second, there is always the possibility of "fresh" carbon leaching into a sample (e.g., via water-bearing organic materials running through a site), resulting in a C14 date that is younger than the actual age of the specimen. Third, the ratio of ^{14}C to ^{12}C in the atmosphere has not remained constant over time, but has varied, probably due to fluctuations in cosmic ray activity and other sources that produce ^{14}C (e.g., there was a big spike in atmospheric ^{14}C levels during the 1950s due to aboveground nuclear bomb tests). C14 dates thus have to be carefully calibrated against this variation in the ratio of ^{14}C to ^{12}C. Still, these (and other) issues are well-known to laboratories that carry out C14 dating, and methods have been developed to either take care of these issues or recognize that the resulting dates may have problems.

Of more concern for our purposes is the interpretation of archaeological dates and evidence, especially when they appear to conflict with genetic evidence. The earliest date for the presence of humans in a particular area, based on archaeological dating, provides a lower limit for when humans actually entered that area. This is because it is highly unlikely that archaeologists will ever find the one site associated with the initial occupation of an area; it is more likely to find sites after humans have had a chance to get established and increase their population size, thereby increasing the number of sites they leave behind. So, we should not be surprised that further archaeological investigations tend to push back the age of the first human presence in an area, nor should we be surprised if genetic dating for a migration event is older than the archaeological dates. Another potential concern with archaeological evidence is the bias in terms of what is actually preserved at a site, as well as the extent to which it is even possible to find the relevant sites in a particular region. Stones, bones, and shells tend to preserve much better than wood and plants, and, moreover, preservation tends to be much better in colder, less humid climates than in warmer, more humid regions. So, the fact that we don't find any sites in a particular region at a particular time may indicate that humans were not present then, or it may be that the climate at that time did not permit preservation of remains; absence of evidence is not the same as evidence of absence.

Turning now to linguistics, there is a natural tendency to expect that the genetic and linguistic relationships of populations should be correlated. After all, if we start with the simple model of a single ancestral population giving rise to two descendant populations, then over time we expect both the genes and the languages of the descendant populations to diverge. Moreover, if two formerly separated populations come into contact, then there is the opportunity for exchange of genes as well as for aspects of one language to show up in the other. For example, some southern African Bantu-speaking groups have genes from Khoisan-speaking groups as well as some "click" consonants that are characteristic features of Khoisan languages, so both the genes and the languages of these Bantu-speaking groups show evidence of contact with Khoisan-speaking groups. This expectation that genes and languages should be correlated is so strongly entrenched that even the great man himself, Charles Darwin, wrote in his 1859 magnum opus, On the Origin of Species, that "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world." In other words, if you want to know how languages are related, first figure out how the populations who speak those languages are related and then use that as your language classification—certainly a strong statement in favor of a close correlation between genes and languages!

However, Darwin notwithstanding, there are also good reasons to expect that the genetic and linguistic relationships of populations would *not* be correlated. Languages tend to change quite rapidly (compare the way you and your peers speak and the words you use to how your parents or grandparents speak!).

The features that linguists use to reconstruct how languages are related—described in more detail below—may change too fast, or change too often in parallel in different languages, to provide reliable indications of distant relationships. It is generally thought to be impossible to demonstrate any relationship among languages that diverged more than about 10,000 years ago or so; that is, two languages that diverged 10,000 years ago will appear as different as two languages that diverged 20,000 years ago or 40,000 years ago (although this assertion appears to be based mostly on opinion rather than on any firm foundation in fact). Moreover, when two different populations come into contact, there can obviously be exchange of genes without any impact on languages or vice versa: nowadays, the English language is much more widespread than the genes of native English-speaking people, with English words creeping into all sorts of different languages. So, one should not automatically expect to find similarities between the genetic and linguistic relationships of populations.

Languages have a number of different features that can be studied to learn about language relationships—just as there are a variety of different kinds of genetic markers that can be studied to learn about genetic relationships, as we saw back in Chapter 7. The most widely studied features are **cognates**, which are words in two (or more) different languages that are derived from the same word in an ancestral language. For example, both the English word "father" and the German equivalent ("Vater") are derived from the Latin "pater." The traditional approach to historical linguistics relies on the identification of cognates to reconstruct ancestral forms of languages (protolanguages) and requires intensive study of the languages in question, making it difficult for nonspecialists to understand or verify claimed relationships. Nowadays, it is becoming more common to apply phylogenetic methods to reconstruct language relationships on the basis of cognates—one can, for example, construct a distance measure based on the proportion of shared cognates between two languages, calculate the linguistic distance among all pairs of languages in your study, and then feed this distance matrix into a tree-building program; we'll see an example of this approach later in this chapter. Using phylogenetic methods to figure out language relationships is not without critics—recall from our discussion of such methods in Chapters 10 and 11 that all phylogenetic methods assume some underlying model as to how the genetic characteristics evolve, and the extent to which languages change according to such models is a continued source of debate—but at least it is easier for the nonspecialist (like me) to understand the results of phylogenetic methods (although there still remains the issue of how to assess the reliability of the linguistic data used in such analyses). Moreover, it is easier to investigate questions such as how strongly the data support particular language groupings, what aspects of the data support (or do not support) particular language groupings, and so forth.

Cognates (words) are not the only source of information about linguistic relationships; there are also structural features of languages, which include phonology (the various sounds that a language uses) and grammar. Many grammatical features vary among languages—for example, in some languages the verb comes in between the subject and the object of a clause, while in others the verb comes after the subject and the object. This variation can be treated and analyzed just like cognates (or genetic markers), for example, constructing distance matrices based on shared/different structural features among a set of languages and then applying phylogenetic methods. This approach is not without controversy—not only do some linguists frown on phylogenetic methods in general, but there are those who think that structural features are not suitable for reconstructing language relationships (i.e., the presence of the same structural feature in different languages can easily arise independently and hence is not a good indication that the languages are related). Still, the value of structural linguistic data for inferring language relationships is a question that should be addressed empirically; in one notable example, the linguist Michael Dunn and colleagues used structural data to demonstrate relationships among a set of Papuan languages from island Melanesia for which the cognates were too divergent to reconstruct any such relationships (Dunn et al. 2005), so it would seem that there is some merit in investigating structural data.

Understanding how languages are related, therefore, can provide insights into how the populations that speak those languages are related. Moreover, there is ample evidence for the expansion of various language families that were probably associated with human migrations and dispersals, and we shall see examples later in this chapter (although keep in mind that languages can spread by cultural transmission, so the spread of a language does not necessarily entail the migration of people speaking that language). Currently, the main limitation of linguistic data for the purposes of this book is that it is not clear how far back in time one can reliably reconstruct language relationships. There is some hope that by focusing on features of languages that are extremely stable, such as some grammatical features or some words that don't change very much, it might be possible to demonstrate relationships among languages that diverged more than 10,000 years ago—much as by focusing on slowly evolving regions of the genome (such as histone genes), we can deduce more ancient genetic

relationships than by focusing on rapidly evolving regions (such as STR loci). But the jury is still out as to how far back in time one will be able to infer relationships among languages.

OUT OF AFRICA—HOW MANY TIMES, WHEN, AND WHICH WAY DID THEY GO?

Given that modern humans arose in Africa, the natural question that then arises is the history of the dispersal(s) of modern humans from Africa: namely, how many were there, when did they occur, and which way did they go? There have been several different ideas put forth concerning the number of dispersals of modern humans from Africa. Way back in 1994, the anthropologists Marta Lahr and Rob Foley wrote an influential paper (Lahr and Foley 1994) suggesting that there were multiple dispersals of modern humans from Africa throughout the Pleistocene, beginning about 100,000 years ago (Figure 16.2). Based largely on fossil and archaeological evidence (as genetic evidence was still rather scanty back then), they argued that the earliest dispersal from Africa went by a southern route, along the coast of India, eventually reaching Sahul (the landmass consisting of Australia and New Guinea, which were connected until rising sea levels separated them about 8000 years ago) around 40,000–50,000 years ago. As sea levels were much lower then, much of the potential fossil/archaeological evidence

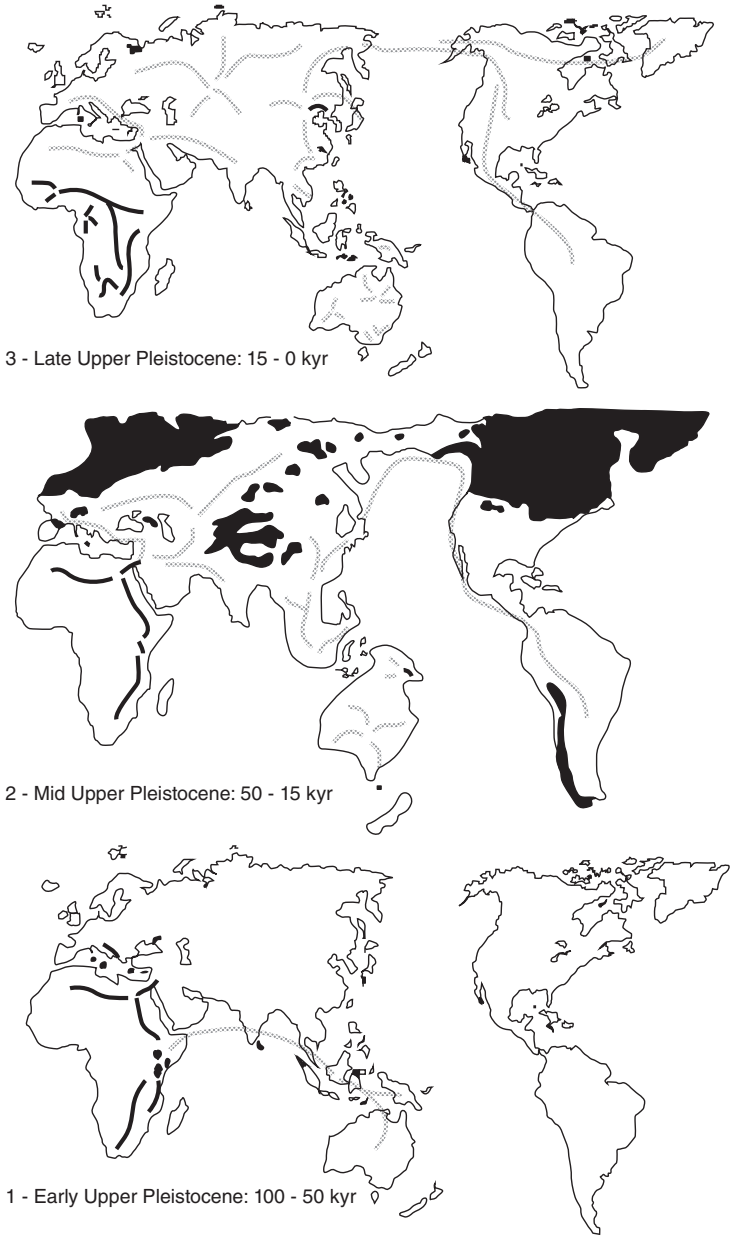


FIGURE 16.2

Multiple dispersals of modern humans from Africa, inferred largely from archaeological and fossil evidence. Black areas show ice sheets during the middle Pleistocene, formally defined as the period from 126 to 781 kya. Reprinted with permission from Lahr, M.M., and Foley, R., "Multiple dispersals and modern human origins," *Evolutionary Anthropology* 3:48, 1994.

of this putative "early southern dispersal," as it came to be known, would now be underwater. Subsequent dispersals would have incorporated or replaced populations descended from this proposed early southern dispersal, thereby erasing any genetic signature of this event, except perhaps in certain more isolated populations. These have been suggested to include aboriginal Australians, Papuans of New Guinea, Andaman Islanders, and so-called "Negrito" groups; these latter include some populations from the Philippines and Malaysia characterized by smaller stature, darker skin pigmentation, and frizzier hair than neighboring groups. Initially, some anthropologists thought—incorrectly—that the Negrito groups were directly related to African Pygmies (whence the name, which means "little black people"), and it has long been supposed (albeit based on very little actual evidence) that the Negrito groups are descended from the earliest migration of humans to the region. We'll come back to this question later in this section.

After the early southern dispersal, subsequent dispersals from Africa to Asia may have also gone by this southern route, while others would have gone north of the Himalayas; dispersals to Europe would also have occurred after this early southern dispersal. The picture that thus arises is one in which our ancestors are more or less continually on the move across the Old World.

However, as mtDNA evidence began to accumulate, the picture seemed to shift in favor of a single major dispersal of modern humans from Africa. Practically all mtDNA types outside of Africa fall into just one of two major haplogroups, called M and N, with estimated ages of around 50,000–70,000 years ago (Figure 16.3). MtDNA evidence based on mismatch distributions (discussed in Chapter 12) also indicates a strong increase in human population size around this time, possibly associated with the expansion of modern humans out of Africa. Haplogroups M and N are both derived from haplogroup L3, which is just one of the numerous mtDNA lineages found in Africa (Figure 16.3). It seems highly unlikely that multiple dispersals of modern humans from Africa would bring only haplogroup L3 out of Africa; a single dispersal seems much more consistent with the mtDNA data. Moreover, many mtDNA lineages branch directly from the ancestral nodes (Figure 16.4), whereas with multiple dispersals to different parts of the Old World, one would expect to find different mtDNA clade structures (i.e., different founder types) corresponding to the different dispersals. These patterns led to the proposal of a single major dispersal of modern human mtDNA from Africa (and, by implication, of modern humans, since mtDNAs do not disperse by themselves!). Corresponding results from the Y chromosome, plus some analyses coming from genome-wide data (e.g., the serial bottleneck model discussed in Chapter 14) also were viewed as supporting a single major dispersal of modern humans from Africa.

However, claiming that the genetic results "support" the single dispersal model is not quite the same as stating that other explanations for the data have actually been tested and found wanting. In particular, the mtDNA results were not subjected to any sort of tests to see which models were—and were not—convincingly supported by the data. Instead, support for the single dispersal model was based on assertions that the observations mentioned previously are more likely under a single dispersal model than under a model of multiple dispersals. While this may very well be correct, it is not a very satisfactory state of affairs.

Single versus multiple dispersal models were subsequently formally tested in a study by Andreas Wollstein, Manfred Kayser, and myself (Wollstein et al. 2010) using genome-wide SNP data (these data were already mentioned back in Chapter 11). Data were obtained for ~800,000 SNPs from populations from Borneo, highland Papua New Guinea, Fiji, and several Polynesian islands (Fiji and Polynesia were sampled for examining hypotheses about the colonization of the Pacific, discussed later in this chapter) and included publicly available data (from the HapMap project mentioned in Chapter 9) from Yoruba, Europeans, and Han Chinese. The approach taken was to model three different population histories, depicted in Figure 16.5, and determine how much support each model received from the data. This was done by carrying out simulations and seeing how many times each model produced summary statistics that were closest to the observed summary statistics for the data.

The three models in Figure 16.5, drawn as branching diagrams, can be interpreted as follows. The leftmost depicts a single dispersal of modern humans from Africa, followed by a single dispersal to Asia and New Guinea, and receives moderate support from the data, with a support value of 0.24 (support values range from 0 to 1; the higher the support value, the better the model fits the data). The middle diagram depicts a single dispersal from Africa, followed by an early dispersal to New Guinea and a subsequent dispersal from this non-African source population to Asia (and Europe), and receives the strongest support from the data (support value = 0.74). The rightmost diagram depicts the "classical" model of multiple dispersals from Africa, with an early dispersal to New Guinea, followed by a subsequent dispersal to Europe and Asia, and receives hardly any support at all from the data (support value = 0.02). So, the conclusion from this analysis is that there was a single dispersal of modern humans from Africa, followed by multiple dispersals from this non-African source population, with the earliest dispersal along a southern route to New Guinea, followed by subsequent dispersals to Europe and Asia.

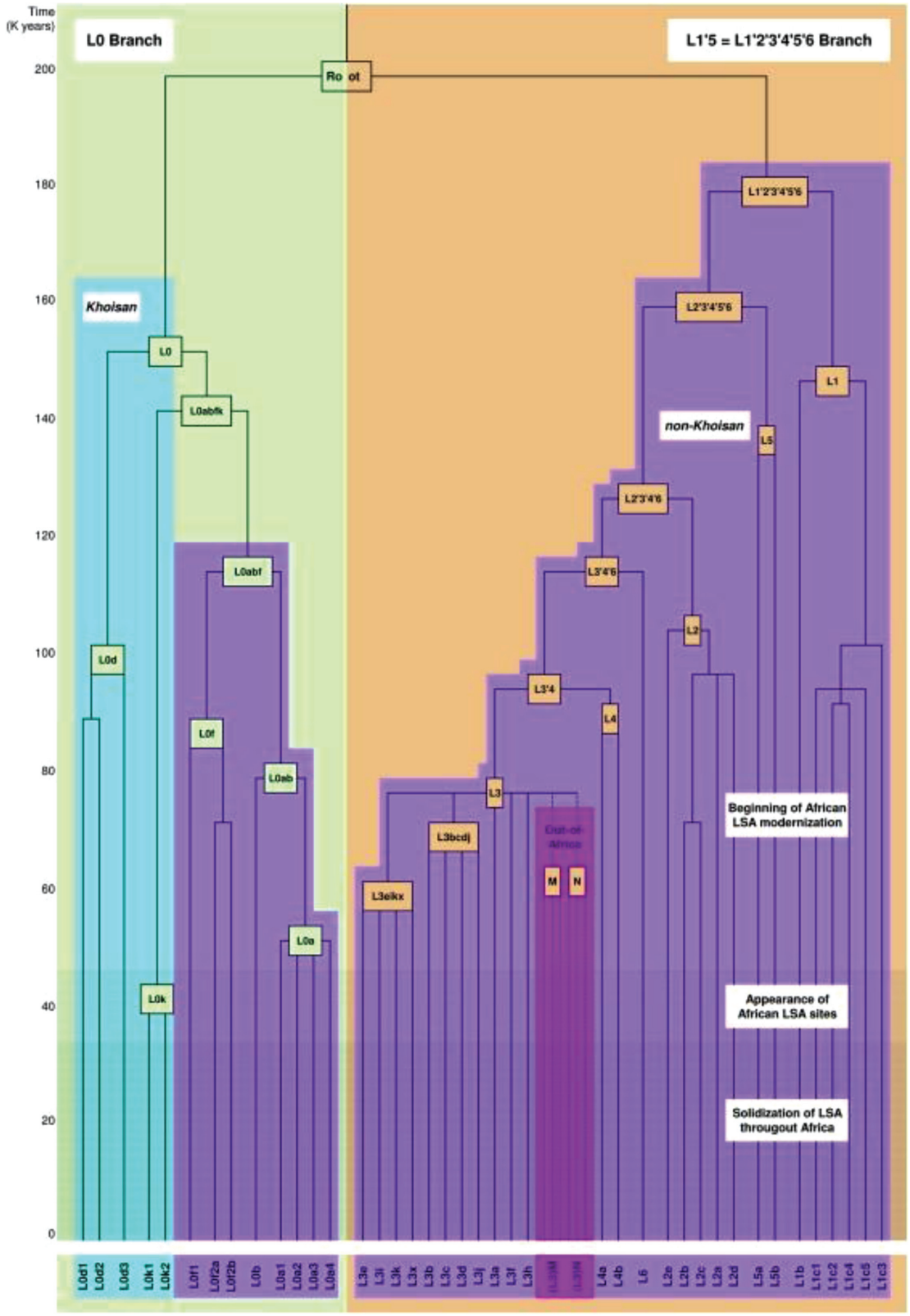


FIGURE 16.3

Human mtDNA phylogeny, emphasizing the diversity in Africa compared to the much more limited diversity outside Africa. Compare this view of the mtDNA phylogeny, showing all the L subhaplogroups in Africa, to that in Figure 9.4, which shows the major "named" haplogroups and the Eurasian bias behind the designation of such haplogroups. Reprinted with permission from Behar, D.M., et al., "The dawn of human matrilineal diversity," *American Journal of Human Genetics* 82:1130, 2008.

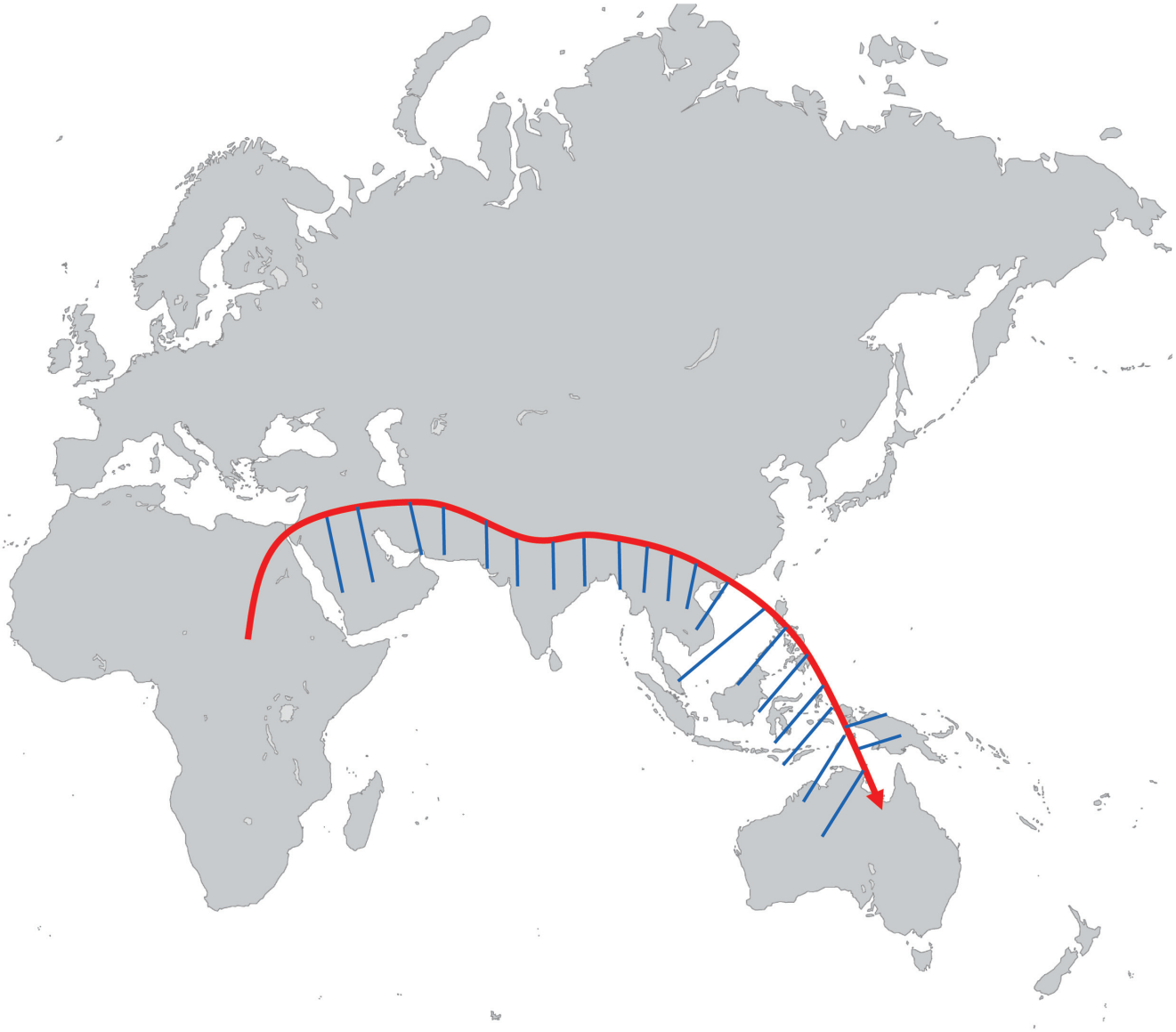


FIGURE 16.4

Radiation of mtDNA types (blue lines) from one migrating mtDNA lineage from Africa (red arrow); this sort of radiation has been argued to support a single, rapid migration of people via a southern route to Southeast Asia and Oceania (e.g., Macaulay et al. 2005).

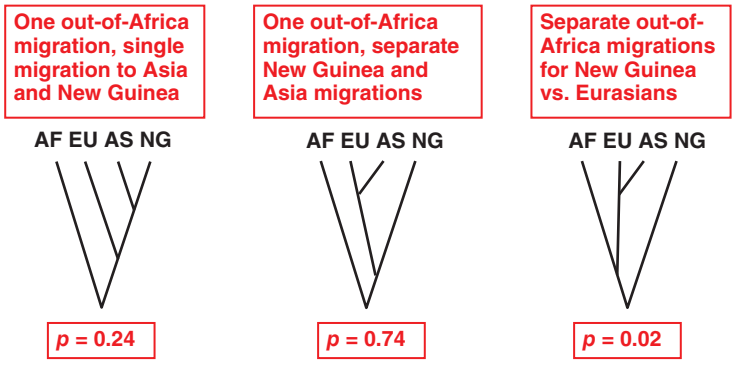


FIGURE 16.5

Three models for the dispersal of modern humans from Africa to New Guinea and associated support values, based on an analysis of genome-wide SNP data. Modified with permission from Wollstein, A., et al., "Demographic history of Oceania inferred from genome-wide data," *Current Biology* 20:1983, 2010.

To be sure, numerous caveats are in order. The sampling of populations is pretty limited—maybe if more or different populations were sampled, a different answer would be obtained. Ascertainment bias—that is, how SNPs were chosen to be included on the commercial SNP chip used in the analysis—is always an important issue with SNP chip data, although Andreas Wollstein, the graduate student who carried out most of the analyses, came up with a very clever way to deal with this problem. Briefly, Andreas compared the difference in summary statistics between SNP chip data and complete sequence data (which is free from any ascertainment bias) for the HapMap populations (for which both types of data are available) and incorporated this into the modeling. Most importantly, the approach used in this study is highly dependent on the simulation approach and the associated assumptions; if the assumptions do not hold (e.g., the model assumes no migration following population divergence, when in fact as discussed in the "Into even more remote lands: the colonization of Polynesia" section, there probably was migration), then the support values may not be meaningful. It should also be pointed out that while the "classic" model of multiple dispersals from Africa is ruled out by this analysis, the best-fitting model of a single dispersal from Africa, followed by multiple dispersals to New Guinea/Asia is not significantly better than the single dispersal from Africa, single dispersal to New Guinea/Asia model. The fact that this latter model (the leftmost model in Figure 16.5) receives appreciable support may be explained if the scenario in the middle model in Figure 16.5 is actually correct, but there was subsequent gene flow from Asia to New Guinea, as seems likely (although this is conjecture and needs further investigation).

However, it turns out that the scenario of a single dispersal of modern humans from Africa, followed by multiple dispersals from this non-African source population, also receives strong support from the signals of archaic human ancestry in modern humans. Recall from the previous chapter on ancient DNA that all non-Africans carry approximately the same signal of Neandertal DNA. The simplest explanation for this signal is that there was a single dispersal of modern humans from Africa that then interbred with Neandertals before dispersing to the rest of the Old World (and, ultimately, to the rest of the globe). Where this interbreeding between modern humans and Neandertals occurred is unknown, but a likely guess is somewhere in the Middle East, as this is the region closest to the possible exit points from Africa where both Neandertal and early modern human fossils have been found. Anyway, wherever the Neandertal interbreeding took place, the fact that a shared signal of Neandertal ancestry is found in all non-Africans would seem to indicate a single dispersal of modern humans from Africa. The very latest evidence (Vernot et al. 2016) infers a total of three episodes of interbreeding between Neandertals and the ancestors of various human populations: one which is shared by all non-Africans; an additional pulse shared by Europeans, South Asians, and East Asians, and a further pulse that is exclusive to East Asians. In hindsight, these multiple events of interbreeding are not surprising–after all, if they could do it once, for sure they could do it more than once. Moreover, note that these results also support separate dispersals of New Guineans (with one episode of Neandertal admixture in their history, the one shared by all non-Africans) and East Asians (with three episodes of Neandertal admixture in their history).

And what about the signal of interbreeding with Denisovans? Although in the previous chapter it was stated that the genetic signal from Denisova was present only in Melanesians, there is actually more to it than that. When the Denisova genome sequence was published, only a small number of populations from southeast Asia and Oceania had been analyzed for any genetic contribution from Denisova (Figure 16.6a)—after all, it was hardly expected that this is where the genetic signal from Denisova would turn up! Indeed, a subsequent study led by myself and David Reich (Reich et al. 2011) found signals of Denisova interbreeding in many additional populations from Southeast Asia and Oceania (Figure 16.6b). However, only populations in the eastern part of island Southeast Asia (e.g., eastern Indonesia and the Philippines) and Oceania (Australia, New Guinea, and islands to the east, including Polynesia) showed a significant signal of Denisova interbreeding; no populations in western Indonesia or anywhere on the Asian mainland showed any significant signals. Importantly, populations that do *not* show any signal of Denisova interbreeding include two groups thought to be related to the groups that do show signs of Denisova interbreeding: these two groups are the Jehai from Malaysia, thought to be related to Philippine Negrito groups (such as the Mamanwa, who do show a signal of Denisova interbreeding), and Andamanese Islanders, thought to be related to Negrito groups and (possibly) to aboriginal Australians and Papuans.

So what does all of this information from signals of Denisova interbreeding (or lack thereof) in various populations tell us about human dispersals? A model of population history was developed and tested that provides the best fit to the data, using various statistics such as the f_4 statistics described in Chapter 12; a schematic version of the model is shown in Figure 16.7. An important caveat to keep in mind is that while the resulting model does provide a better fit to the data than any other model, we do not know whether it is significantly better than other models. This is because the statistics used to assess the fit of the data to the model (f_4 statistics and the like) are

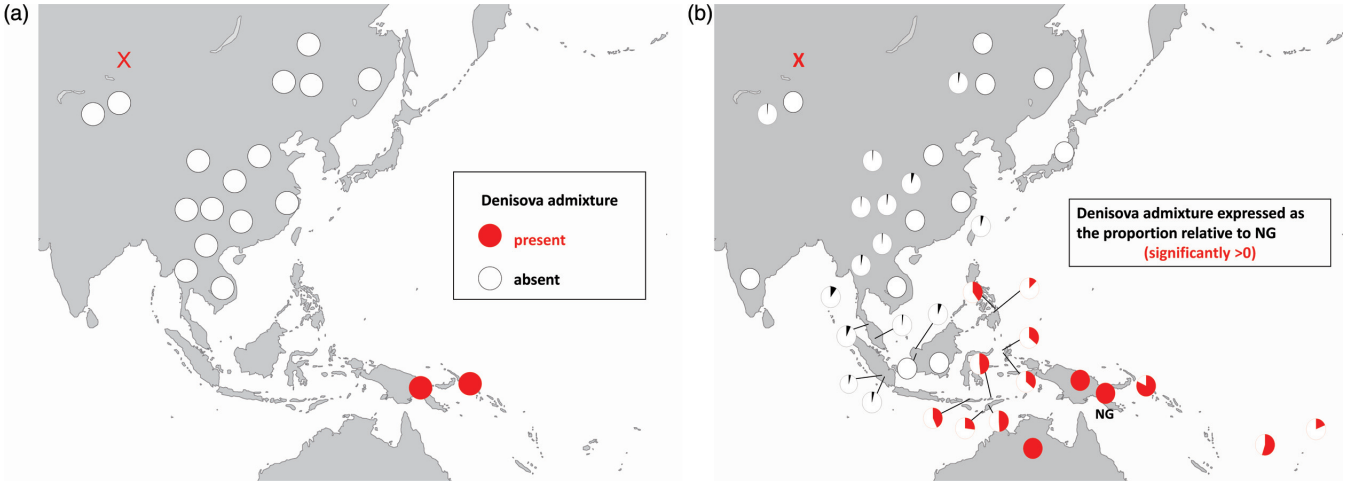


FIGURE 16.6

(a) Populations from Asia and Oceania that were analyzed for Denisova admixture in the Denisova genome study. The red X marks the position of Denisova Cave, and red circles indicate populations with significant amounts of Denisova ancestry. (b) Estimated Denisova admixture in additional populations from Southeast Asia and Oceania, with the pie charts indicating the amount of Denisova ancestry relative to that in New Guinea (black, Denisova ancestry not significantly different from zero; red, Denisova ancestry significantly greater than zero). (a) Modified with permission from Reich, D., et al., "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature* 468:1053, 2010. (b) Modified with permission from Reich, D., et al., "Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania," *American Journal of Human Genetics* 89:516, 2011.

all correlated with one another in a very complicated fashion, which prevents any formal test of statistical significance. Anyway, the best-fitting model indicates that after modern humans left Africa, the first dispersal probably was by a southern route, as existing groups that are descended from this dispersal include the Andamanese and the Jehai (a Malaysian Negrito group). After the ancestors of the Andamanese and the Jehai diverged, there was a single interbreeding event with Denisovans, followed by the divergence of ancestors of the Mamanwa (a Philippine Negrito group), and then the divergence of aboriginal Australians and Papuans. Later dispersal events brought the ancestors of all other east and Southeast Asian groups (such as Han Chinese, western Indonesians, etc.). The other groups in the area who exhibit signals of Denisova interbreeding (eastern Indonesians, Fijians, and Polynesians, cf. Figure 16.6b) all share recent ancestry/contact with Papuans, and their signals of Denisova interbreeding most likely came about secondarily via this shared ancestry/contact with Papuans and not directly from interbreeding with Denisovans. So, the overall picture from the signals of Denisova interbreeding in some populations (and the lack of the signal in others) fits very nicely with the results discussed previously, based on modeling of genome-wide SNP data in New Guinea and other populations: namely, an early dispersal, most likely via a southern route, to Southeast Asia and Oceania.

A related question that arises from this work is: where did the interbreeding between modern humans and Denisovans take place? To be sure, given the propensity for modern humans to move around, it is risky to try to make inferences about where events took place in the past based on their signals in current populations. The populations in a particular region today may not be at all representative genetically of the populations that were there thousands and thousands of years ago. Moreover, even though all of the populations that exhibit signals of Denisova admixture are several thousand kilometers away from Denisova Cave, perhaps the ancestors of the modern human groups did pass by the vicinity of the cave and that is when the interbreeding took place. Still, the fact that the genome-wide data indicate that the Andamanese, Jehai, Mamanwa, Papuans, and aboriginal Australians are all descended from the same migration, but the first two lack any signal of Denisova interbreeding while the latter three all have it does suggest that the Denisova interbreeding took place somewhere in the vicinity of island Southeast Asia (cf. Figure 16.7). And, if this is indeed the case, then it tells us something not only about human dispersals but also about the capabilities of Denisovans: if Denisovans were in island Southeast Asia, then they were spread across a wider geographic and climatically variable area (i.e., from the taiga forests of southern Siberia to the tropical jungles of Southeast Asia) than any other hominin (except us).

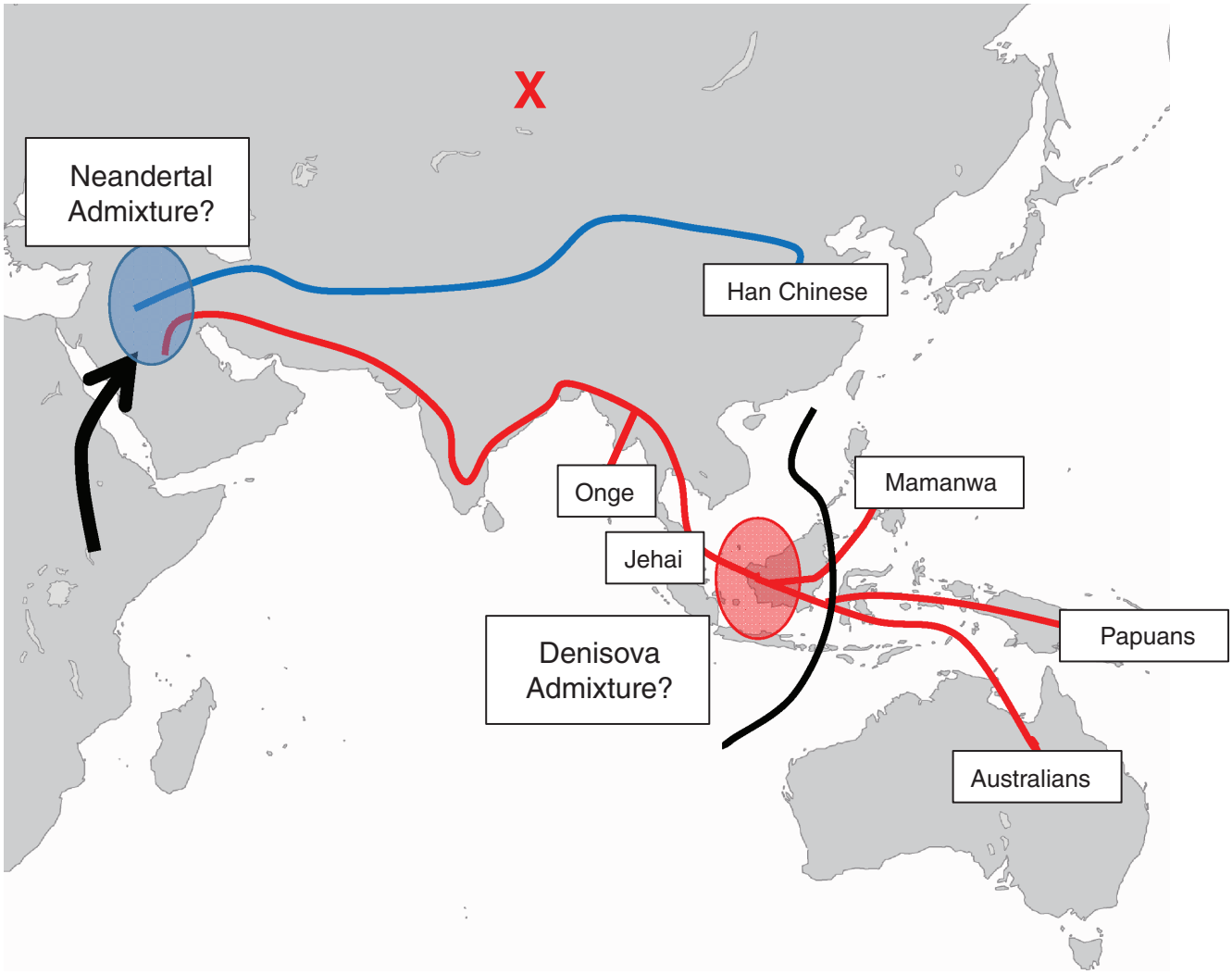


FIGURE 16.7

Schematic version of the best-fitting model for the dispersals of modern humans, suggested by the signals of archaic human ancestry in modern humans. The red line shows a hypothetical path for the early southern dispersal, with Andamanese and Jehai branching off first, followed by Denisova admixture and then branching off of Mamanwa, Australians, and New Guineans. The blue line represents additional dispersal(s) leading to Han Chinese (and all other Asian populations), while the black line is the boundary between populations with and without Denisova admixture. Reprinted with permission from Stoneking, M., "Archaic genomes and the peopling of South Asia," in G.R. Schug and S.R. Walimbe (editors), *A Companion to South Asia in the Past*, John Wiley & Sons, Inc.: New York, 2016.

So, to summarize, analyses of genome-wide SNP data in modern humans for inferences about demographic history as well as for signals of interbreeding with archaic humans (Neandertals and Denisovans) both give rise to the same picture (Figure 16.7): namely, a single major dispersal of modern humans from Africa, followed by interbreeding with Neandertals, followed by an early southern route dispersal (with interbreeding with Denisovans somewhere along the way), followed by additional dispersals from this non-African source population to Europe and Asia (along with additional episodes of Neandertal admixture).

However, as is often the case, additional subsequent work has shown that the neat and tidy version of events depicted in Figure 16.7 is not quite correct; one often gets the feeling in this business that the more research we do, the less we end up knowing, so maybe we should quit while we are ahead! It now appears that there is a signal of Denisovan ancestry that is widespread across East Asia and the New World (Qin and Stoneking 2015). To be sure, this Denisovan

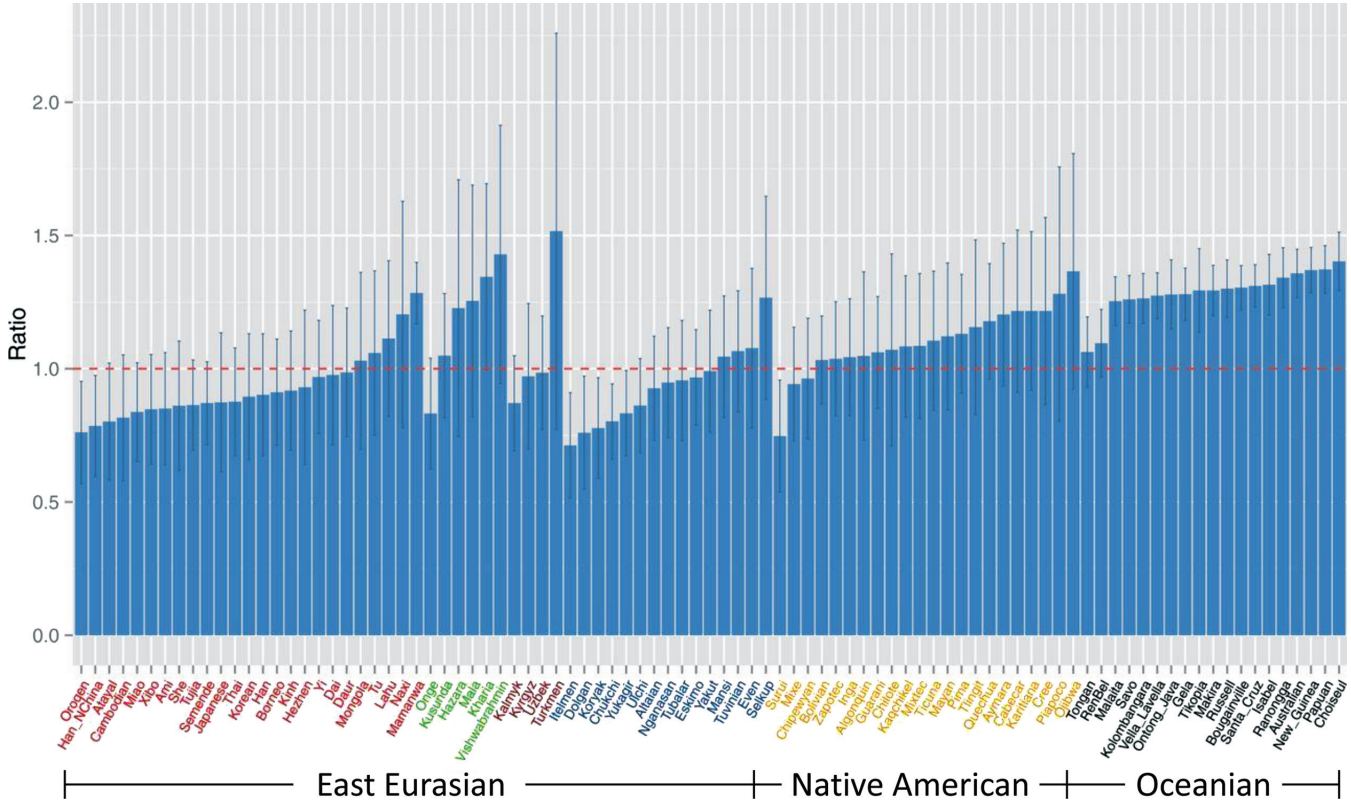


FIGURE 16.8

Evidence for widespread Denisovan ancestry in East Eurasian and Native American populations, as well as in Oceanians. Plotted is the ratio of estimated Denisovan ancestry to estimated Neandertal ancestry; values in excess of 1 (dashed red line) are evidence for Denisovan ancestry. Reprinted with permission from Qin, P., and Stoneking, M., "Denisovan ancestry in East Eurasian and Native American populations," *Molecular Biology and Evolution* 32:2665, 2015.

ancestry cannot be detected in every East Asian or New World population, but it is present at (barely) detectable levels, around 0.2% (as compared to the 4–6% Denisovan ancestry in Oceania) in enough populations (Figure 16.8) that it makes more sense to think that it was present in every population, and by genetic drift decreased to undetectable levels in some populations, rather than proposing that the ancestors of only some East Asian and New World populations interbred with Denisovans. Moreover, the Denisovan ancestry in East Asia and the New World seems to be related to that in Oceania. Note that the Denisovan ancestry in the New World can be readily explained if admixture with Denisovans occurred in the East Asian ancestors of New World populations prior to the colonization of the New World, which is what one would expect given the evidence for initial colonization of the New World around 15,000 years ago (as discussed in the "Into remote lands: the colonization of the Americas" section). There are (at least) two scenarios that could account for the Denisovan ancestry in East Asia: first, interbreeding with Denisovans may have occurred specifically in the ancestors of Near Oceanians (cf. Figure 16.7), but before the subsequent divergence of Philippine/Australian/New Guinean populations, there was then a back migration to East Asia that contributed a small amount of Denisovan ancestry to East Asian populations. Or, Denisovan admixture may have occurred elsewhere, contributing Denisovan ancestry at more or less the same level across East Asia and Near Oceania, but subsequent migration(s) of humans (lacking Denisovan ancestry) contributed ancestry to East Asia, thereby "diluting" their Denisovan ancestry, but Near Oceania was not impacted by this additional ancestry and hence maintained higher levels of Denisovan ancestry. Which of these might be correct—or indeed, if the picture is actually even more complicated—is not at all clear as I write this.

And if that isn't already complicated enough, there is recent evidence that the Altai Neandertal—but not any other Neandertal analyzed so far—has a small amount of modern human ancestry (Kuhlwilm et al. 2016). Intriguingly, this modern human ancestry, while related to Africans, seems to be older than

current non-African human populations, dating to at least 100,000 years ago. These results suggest that the ancestors of the Altai Neandertal met up with an early population of modern humans outside of Africa (since so far there is no evidence that Neandertals ever went back into Africa), and this early population of modern humans subsequently went extinct. However, this is not the only explanation, and as I write this there is great uncertainty as to what actually happened in terms of multiple dispersals to and from Africa—as well as great anticipation for what further studies of archaic and early modern human remains will reveal about this phase of our evolutionary history.

INTO REMOTE LANDS: THE COLONIZATION OF THE AMERICAS

Christopher Columbus set sail from Palos de la Frontera, Spain, on August 3, 1492, hoping to find a new and quicker route to the riches of east and south Asia (known as the "Indies"). The voyage took longer than either he or his crew anticipated, and worried that the crew would mutiny, he kept two different sets of records, one showing the true distance they had traveled, the other showing a shorter distance; this latter log he showed the crew in an attempt to convince them that they really hadn't traveled so far. This strategy worked for only a short time, however, and eventually Columbus was forced to promise the crew that they would turn back if land was not sighted within 2 days. Fortunately for Columbus, land was sighted the very next day, and on October 12, 1492, they made landfall on an island in what is now the Bahamas. Although this date is widely celebrated as the anniversary of the discovery of the New World, in reality what Columbus "discovered" was a land already populated by millions of people whose ancestors had been there for thousands of years, whom he called Indians (thinking that he had indeed reached the Indies).

When and from where did people first reach the Americas? And, how many waves of migration were there to the New World before Europeans managed to arrive on the scene? These are the questions we will address in this section. Before delving into what genetics has to say about these questions, let's first quickly review the archaeological and linguistic evidence. It was pretty clear from the beginning that Native Americans were of East Asian origin, notwithstanding other ideas such as boating from Europe across the Atlantic, or being a "lost" tribe of Israel. However, many anthropologists assumed that people would not have had the wherewithal to get to the Americas until a few thousand years ago and vigorously resisted any suggestions of an earlier colonization of the New World. This view changed with discoveries of distinctive spear and arrow

FIGURE 16.9

Examples of Clovis points, from the Rummells–Maske site in Iowa. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Clovis_Rummells_Maske.jpg).

points in the 1930s at a site near Clovis, New Mexico (Figure 16.9) in association with mammoth remains that were around 10,000 years old. Incidentally, one of the most prominent skeptics-turned-converts was Aleš Hrdlička, curator of physical anthropology at the Smithsonian Museum and founder of the "American Journal of Physical Anthropology" (AJPA); during the years he was editor of AJPA, he discouraged the use of statistical methods in papers published in the journal, declaring that "statistics would be the ruin of physical anthropology" (quoted in Montagu 1944). Other sites dating to around the same time and with the same Clovis-style tools were found across North America, leading to the "Clovis-first" paradigm for the colonization of the Americas. According to Clovis-first, the first people to reach the New World came via a land bridge across the Bering Strait about 11,500 years ago and quickly exterminated the large megafauna such as mammoths, mastodons, giant ground sloths, and so forth (or at least, contributed to their extinction, as some researchers think the megafauna were already stressed by climate change and on their way out when humans arrived on the scene). The archaeological evidence—Clovis sites appearing all across North America within the span of a few hundred years, and the extinction of the megafauna at about the same time—would certainly seem to fit with a picture of humans arriving for the first time in the Americas around 11,500 years ago and finding the megafauna to be easy prey for their relatively sophisticated hunting technology, enabling the Clovis hunters to spread quickly across North America. Note that much of northern North America was covered by ice sheets around this time, but an ice-free corridor is thought to have existed beginning about 14,000 years ago that would have facilitated travel through this region. Alternatively, it has been suggested that humans may have moved along a coastal route, perhaps using boats.

While "Clovis-first" still has its adherents, it is now generally accepted that humans were in the New

World before Clovis. Currently, the oldest securely dated human remains are not bones but rather coprolites (fossil feces—bones are not the only remains ancient humans left behind!) of human origin (confirmed by DNA analysis) from Paisley Caves, Oregon, that have been dated to about 14,000 years ago and are associated with non-Clovis tools (Gilbert et al. 2008). And all the way down in Chile, at a site called Monte Verde, charcoal has been dated to about 12,500 years ago (Dillehay 1989), with a recent study claiming evidence for human occupation at least 14,500 years ago and maybe as much as 18,500 years ago (Dillehay et al. 2015). For a long time, there was controversy about this site—not so much the dating, but rather whether the charcoal came from a hearth, indicating human occupation (as claimed by the archaeologist who excavated the site, Tom Dillehay), or rather was from natural sources such as a lightning strike or wildfire. To his credit, Dillehay opened the site for inspection by other archaeologists, and the general consensus is that the charcoal is indeed from a hearth, and the dating is secure (Meltzer et al. 1997). Other sites for which old dates have been claimed, such as the Meadowcroft Rockshelter in Pennsylvania dated to 16,000–19,000 years ago (Adovasio and Carlisle 1988), have not been subject to any independent inspection/verification and hence are considered questionable.

Clearly, if humans were in southern Chile by 12,500 years ago and came via North America (this being the only reasonable route—an alternative route across the Pacific from Polynesia can be ruled out as humans got to Polynesia less than 3000 years ago, as we shall see in the next section), then they must have entered North America at least a few hundred (and more likely, a thousand or more) years earlier. So, humans were in the Americas by at least 14,000 years ago or so. However, an important limitation of the archaeological evidence to keep in mind is that there is no way of knowing to what extent (if any) the people who lived at these various sites are the ancestors of current Native Americans. That is, it could still be the case that the pre-Clovis sites represent one or more incursions into the Americas that didn't persist, and it was only with Clovis that humans finally colonized the Americas once and for all. Or, it could be that pre-Clovis and Clovis sites reflect multiple migrations that did contribute to the genetic ancestry of current Native Americans. Or, it could be that there was a single pre-Clovis migration from which current Native Americans are descended, and Clovis was an indigenous development. As we shall see below, genetic evidence may help sort this out—although, as we shall see, genetic evidence also has limitations.

The linguistic picture for the Americas is as follows: there is one group of about 40 related languages called Na-Dene, spoken by populations that inhabit northern North America or migrated recently from there (such as Apaches and Navajos in the southwestern United States); there is another group of about 15 related languages called Eskimo-Aleut, spoken in the Aleutian Islands, the North American Arctic, Greenland, and the Chukchi Peninsula of Siberia; and then there are all of the other languages of the Americas (an estimated 1500 or so at the time of European contact, of which maybe half still survive), for which there is no consensus as to how they are related. At one extreme, the linguist Joseph Greenberg lumped all of these into one language family called Amerind—thus, three major language families in the Americas (Figure 16.10, left). And based on this conclusion, Greenberg (and others) proposed three waves of migration to the New World, with the oldest bringing the Amerind language family and more recent migrations bringing the Na-Dene and Eskimo-Aleut families. But very few linguists who work on the languages of the Americas find the evidence for a single Amerind language family compelling; instead, they would group the languages into several different families for which no further relationships are demonstrable (e.g., Figure 16.10, right). And an important—and unresolved—question is then, do these different language families reflect multiple migrations to the New World, or do they reflect a single migration and differentiation within the New World that was so long ago that relationships among the various language families can no longer be identified? None of the language families of the Americas show any demonstrable relationship to languages in Siberia or Asia, with the exception of Na-Dene languages, which can be related to Ket, a language currently spoken in the vicinity of the Yenisei River in Siberia (and probably more widespread in earlier times). But this lack of any demonstrable relationships between Asian and New World languages is not particularly surprising, because as mentioned previously, it is thought that languages change too rapidly to retain any signal of a relationship that goes back more than about 10,000 years or so. Moreover, as we shall see when we discuss the genetic evidence, it is quite likely that none of the existing populations in Siberia are directly descended from the ancestral population(s) that colonized the New World, so we should not expect to detect any relationships in their languages.

Keeping in mind the archaeological evidence for initial colonization of the Americas around 14,000–15,000 years ago, and the linguistic evidence for a minimum of three major language families in the Americas that may reflect three (or more) migrations, let's now turn to the genetic evidence. First we'll look what genetics has to say about when the Americas were colonized, then we'll look into the number of migrations. The first detailed DNA analyses, in the early 1990s, involved mtDNA variation and quickly established that

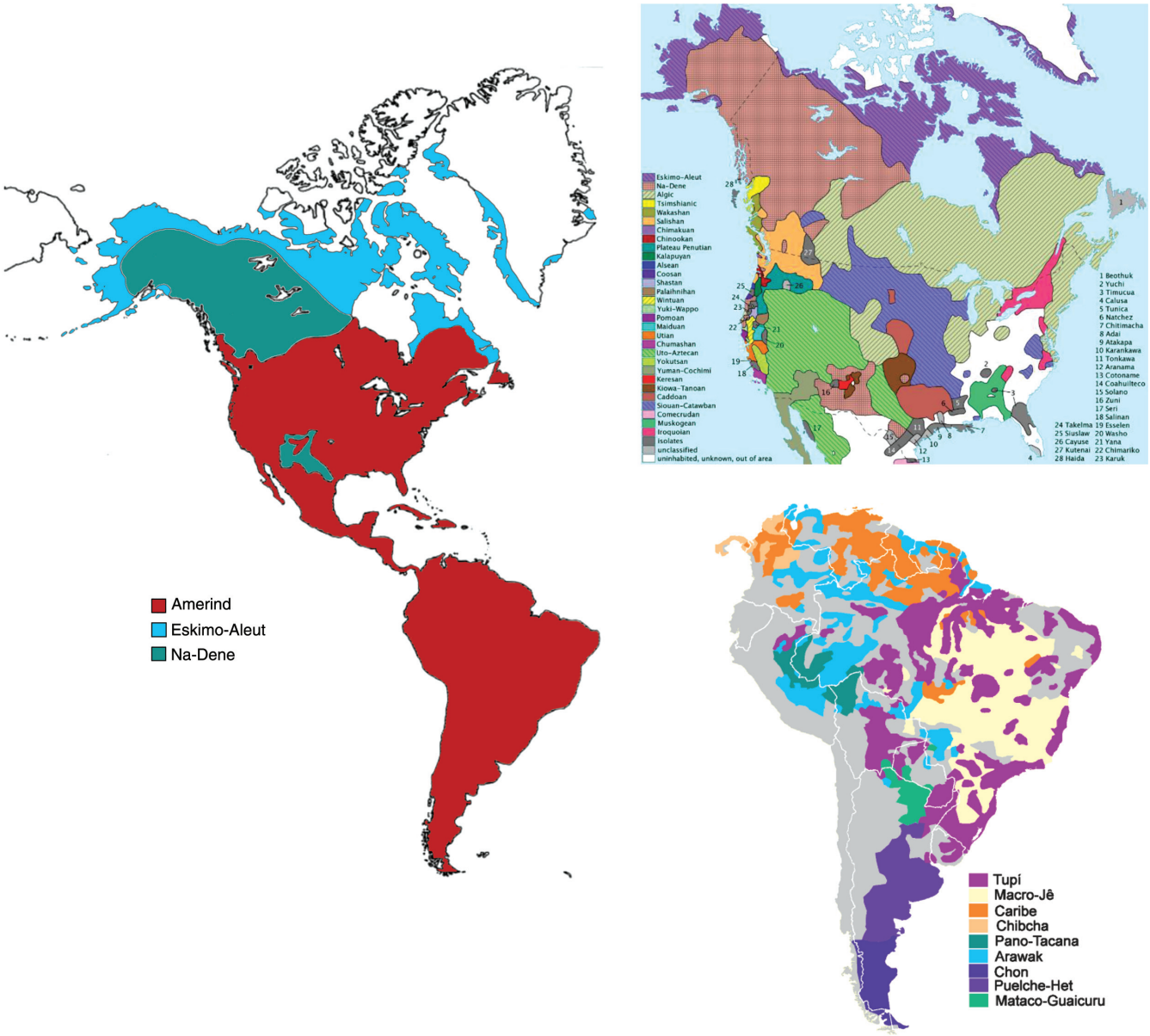


FIGURE 16.10

Two views of the relationships of Native American languages. Left, map of three major language families. Right, map of numerous major language families. Source for the maps on the right: Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Langs_N.Amer.png; https://commons.wikimedia.org/wiki/File:SouthAmerican_families.png).

there is reduced mtDNA variation in the New World compared to most other parts of the world, suggesting a bottleneck. Initial studies (Schurr et al. 1990; Torroni et al. 1993) found four major haplogroups, called A, B, C, and D, as these were among the first haplogroups to be so designated—mtDNA haplogroups were labeled rather haphazardly, as they were discovered, and thus do not correspond in any meaningful way to their phylogenetic relationships (fortunately, as discussed back in Chapter 9, geneticists learned their lesson from all the confusion over the mtDNA haplogroup nomenclature, and NRY haplogroup nomenclature does correspond to the NRY phylogeny). These same four haplogroups occur sporadically in East Asia, supporting an East Asian origin for Native Americans, but it is hard to be more specific than that—Mongolians and Altaians appear most similar genetically to Native Americans, but all one can really conclude is that Mongolians and Altaians are the most similar among contemporary populations to the ancestors of Native Americans, who may have been living far away from Mongolia or the Altai at the time of colonization.

Initially, there was much discussion about how many migrations brought these haplogroups to the Americas: one researcher argued for a separate migration for each haplogroup, which does not seem likely; others argued for one migration that brought haplogroups A, C, and D, and a later migration that brought haplogroup B as it seemed to have less variation than the others; still others argued for a single migration, based on the observation that all of these haplogroups are not so common in Asia, so it is unlikely that multiple migrations would have brought only these four haplogroups and not others that are more common in Asia.

With more detailed studies, including many of complete mtDNA genome sequences, the current picture of mtDNA diversity is as follows. There are actually five mtDNA haplogroups represented in the Americas, and these are now called A2, B2, C1, D1, and X2a. When haplogroup X (now X2a) was found in the Americas, it was known elsewhere only from Europe and the Caucasus, which led some to suggest that haplogroup X supported the "Solutrean hypothesis" for Native American origins. According to this hypothesis, people traveled by boats and ice sheets from northern Eurasia to the New World via the Atlantic; archaeological evidence for this hypothesis is based on perceived similarities between the Solutrean tool tradition, which existed in western Europe between about 17,000 and 21,000 years ago, and Clovis tools. However, most archaeologists do not find the claimed similarities of any significance, and moreover mtDNA sequences that are related to haplogroup X sequences in the Americas have been found in Asia, so the Solutrean hypothesis does not have much in favor of it (other than a few stalwart supporters who refuse to let any facts get in the way of a nice story).

Anyway, these five mtDNA haplogroups all have similar amounts of variation, and a Bayesian analysis of complete mtDNA genome sequences indicates that they each began diversifying and expanding at about the same time (Figure 16.11). Note that in Figure 16.11, haplogroup X2a seems to have a somewhat different signal of expansion, but it is present only sporadically in the Americas and at low frequencies, so it

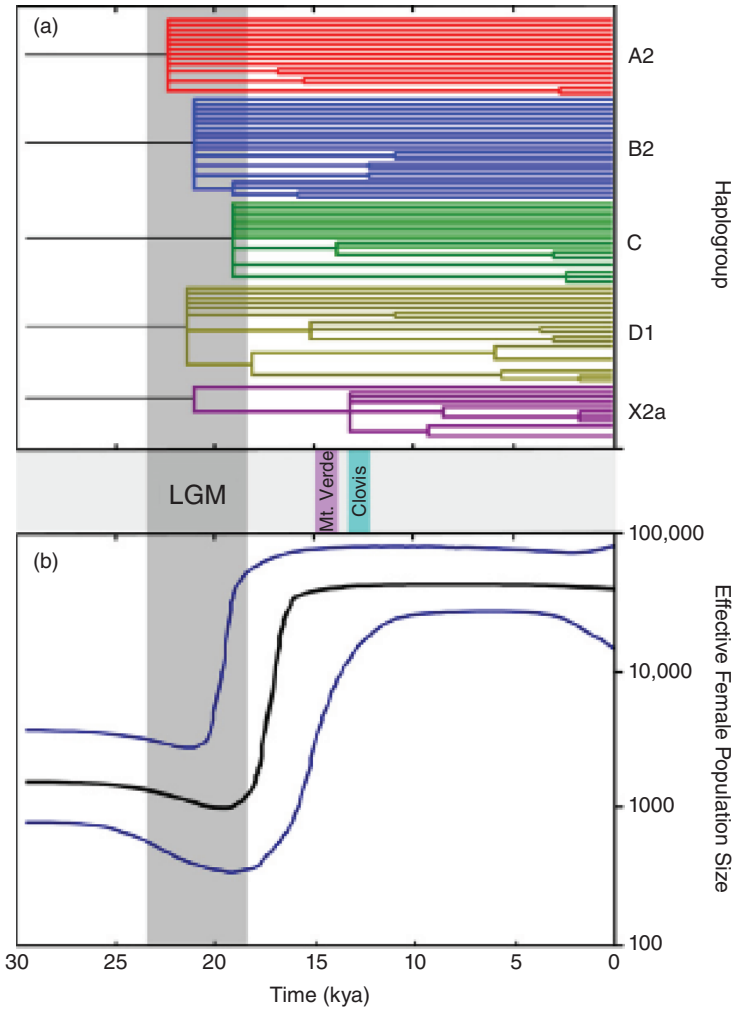


FIGURE 16.11

Phylogenetic relationships and Bayesian skyline analysis (showing population size change over time) for the major New World mtDNA haplogroups. The gray shading indicates the Last Glacial Maximum, and the times associated with the sites at Monte Verde and Clovis are indicated. LGM indicates Last Glacial Maximum. Reprinted with permission from Fagundes, N.J., et al., "Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas," *American Journal of Human Genetics* 82:583, 2008.

is likely that drift has had a bigger impact on this haplogroup than on the other New World haplogroups. These analyses indicated that all of the Native American haplogroups show a signal of differentiation from their respective Asian haplogroups at around 20,000–23,000 years ago, followed by an expansion beginning about 18,000–19,000 years ago that ended about 15,000 years ago. These signals of differentiation and expansion precede the archaeological dates for Monte Verde and for Clovis by a few thousand years, so some geneticists have proposed a "three-stage" or "Beringia-incubation" model to explain this discrepancy (Figure 16.12). According to this model, an ancestral population became separated and isolated from other Asian populations; Beringia is a popular location for where this would have happened, but there is no specific evidence for (or against) this location. This allowed for genetic differentiation of the ancestral population from other Asian populations and the development of the distinctive New World mtDNA haplogroups before the actual colonization of the New World. People would then begin expanding into the New World around 16,000–18,000 years ago, in order to get to Monte Verde by 12,500 years ago. Since the ice-free corridor did not exist at this time, humans presumably moved initially via a coastal route (although the ice-free corridor may have been used at a later time to move north). While this is certainly an attractive model, it does (in my opinion) assume more certainty in the molecular clock dates than I think is warranted. There is sufficient uncertainty surrounding dating with molecular clocks in general and the rate at which the mtDNA molecular clock ticks in particular (as discussed in Chapter 12) that the above dates for diversification and expansion could easily be overestimated by a few thousand years. Moreover, the apparently large genetic differences between Amerindians and current populations in Siberia could reflect population replacement events that occurred in Siberia after people left there for the New World. Ancient DNA evidence suggests a change in population relationships around Lake Baikal between about 7000 and 6000 years ago (Mooder et al. 2006), and a recent analysis of genome-wide SNP data suggests major migrations from southern to northern Siberia during the past 3000 years or so (Pugach et al. 2016). If Siberia was indeed depopulated and then repopulated during this time period, this could explain why current Siberian populations do not show genetic similarities with the presumed ancestral population for Native Americans. And yet another point to keep in mind is that, as discussed in the beginning of this chapter, it is natural to find older dates from genetic than from archaeological evidence, because it is highly unlikely that the archaeologists will ever find the site corresponding to the actual first entry of humans.

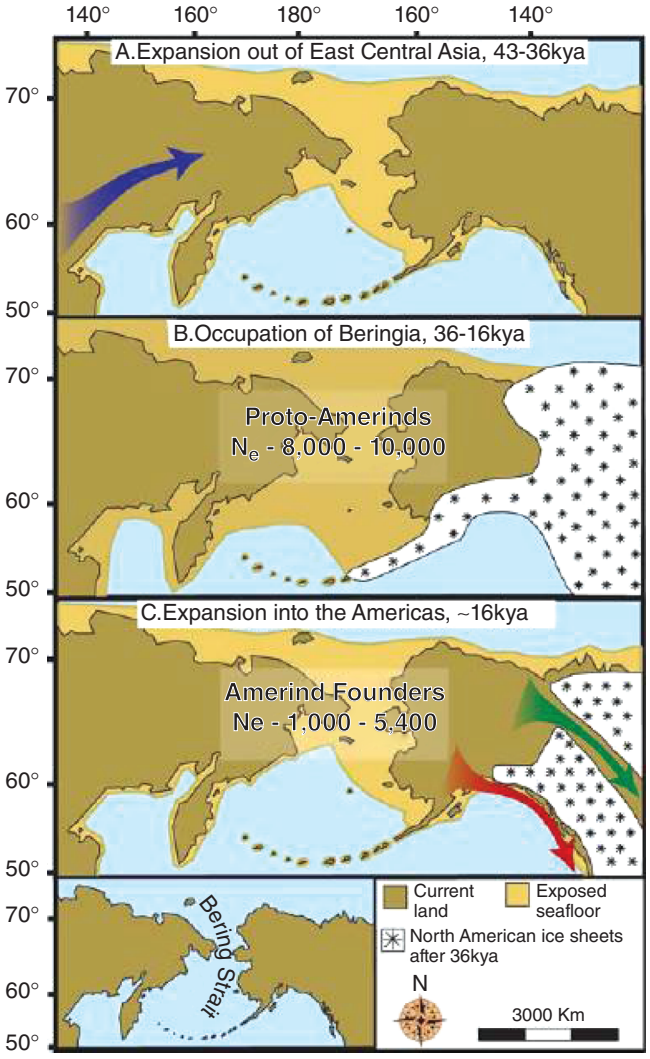


FIGURE 16.12

"Beringian incubation" or "3-stage" colonization model for the origins of New World populations. According to this model, the ancestors of the first people to reach the New World expand out of Asia to Beringia as much as 40,000 years ago, remain isolated in Beringia for up to 24,000 years, and then migrate into the Americas beginning about 16,000 years ago. Modified with permission from Kitchen, A., et al., "A three-stage colonization model for the peopling of the Americas," *PLoS One* 3:e1596, 2008.

Nonetheless, despite some uncertainty over the details, the mtDNA evidence does seem to rather strongly support a single major migration to the New World that is in pretty close agreement with the archaeological evidence. However, there are some substantial differences in the frequencies of the major mtDNA haplogroups (and sublineages) between Eskimo-Aleuts (with mostly A and D lineages), Na-Dene (mostly A in the north vs. A and B in the south,

which might reflect admixture in the southern Na-Dene groups), and other Native American groups (Figure 16.13). Whether these differences reflect isolation and genetic drift within the context of a single migration, or different migrations, is still a matter of debate.

Turning now to the Y chromosome, as with mtDNA there is reduced NRY variation compared to most other populations, with just two NRY haplogroups (C and Q) found in the Americas that are thought to be of prehistoric origins. Several other NRY haplogroups have been found in Native Americans, some of which have been claimed to represent additional founding NRY lineages, but these are found in modern European populations as well and hence most probably represent the all too familiar spreading of genes by European colonists. Incidentally, there is also an appreciable frequency of African NRY and mtDNA lineages in Native Americans, especially in some South American groups, which reflects interactions between people descended from African slaves and the Native American groups. The variation within haplogroups C and Q has been dated by use of associated STR variation to about 10,000 years ago (plus or minus a few thousand years), and so the NRY evidence has been interpreted as favoring a single migration of a limited number of people to the New World somewhere around 10,000–15,000 years ago (Zegura et al. 2004), in good agreement with the mtDNA evidence.

With regard to genome-wide data, to date there have been two major studies of Native American populations. The first (Wang et al. 2007) analyzed nearly 700 autosomal STR loci in more than 500 individuals from 29 Native American populations (including Na-Dene-speaking groups), and the major finding was decreasing genetic diversity and increasing genetic distance with increasing geographic distance from Siberia, in accordance with a serial bottleneck model (as discussed in Chapter 12) for a north to south colonization of the Americas (no surprise there). There was also some indication of genetic differences between populations speaking Na-Dene languages and other Native American populations, as well as more genetic diversity in western than eastern South America—the latter suggesting that people may have moved into South America initially along the west coast and then spread eastward. A subsequent study of the same data used an approximate Bayesian computation approach to analyze various models for the colonization of the Americas and concluded that the best-fitting model (of those examined) was an initial colonization about 13,000 years ago, followed by recurrent gene flow between Asia and the Americas (Ray et al. 2010). This is an important point to keep in mind, as many studies take the easy (or computationally feasible) way out and model gene flow events as discrete events, when continuous (i.e., recurrent) gene flow would seem to be a much more realistic scenario for human migrations. Anyway, the second major study of genome-wide data (Reich et al. 2012) analyzed about 365,000 SNPs in nearly 750 people from 17 Siberian and 52 Native American populations, including one Na-Dene group (Chipewyan) and three Eskimo-Aleut groups. As with the STR data, the SNP data reassuringly show that genetic diversity decreases with increasing distance from the Bering Strait. More intriguingly, the authors used f_4 statistics (discussed in Chapter 12) to identify three distinct streams of gene flow from Asia (Figure 16.14): one that contributed only to the Chipewyan, one that contributed only to Eskimo-Aleut groups, and one that contributed to everyone (including Chipewyan and Eskimo-Aleut groups—so, there was considerable admixture among these different groups). Regrettably, the authors refer to all of the groups who do not speak Na-Dene or Eskimo-Aleut languages as "First Americans" rather than the usual designation of "Amerind"—this sort of terminology should be avoided when discussing genetic results as it can be misused, for example, in disputes over land ownership (i.e., we're the "First Americans" so we were here first and therefore the land belongs to us). Moreover, this study did not try to estimate the timing of these events, nor were more localized sources for the three "Asian" streams of gene flow into the New World identified. It is thus possible that some of the streams of gene flow came from Beringia, and/or the admixture events took place in Beringia rather than the New World, and so the "three stream" scenario could still be compatible with a single major founding migration to the New World. This study also found evidence for back migration from Eskimo-Aleut groups across the Bering Strait to Siberian Eskimo groups, which had previously been hypothesized from linguistic evidence, and which also receives support from mtDNA and NRY evidence.

Ancient DNA analyses have also contributed to our understanding of the peopling of the New World. Such analyses have been technically easier to carry out than ancient DNA analyses of Europeans, as the distinctiveness of Native American mtDNA haplogroups enables identification of contamination from Europeans who have handled the remains or carried out the laboratory analyses. Ancient DNA has verified that the inferred founding mtDNA and NRY lineages are indeed present in the New World at an early age—for example, mtDNA haplogroup B has been found in the 14,000-year-old coprolites from Paisley Cave, and a 10,300-year-old molar from the evocatively named On Your Knees Cave in Alaska was found to have mtDNA haplogroup D as well as a sublineage of NRY haplogroup Q that is widespread in contemporary Native American groups. Moreover, no unexpected mtDNA (or NRY) haplogroups have been

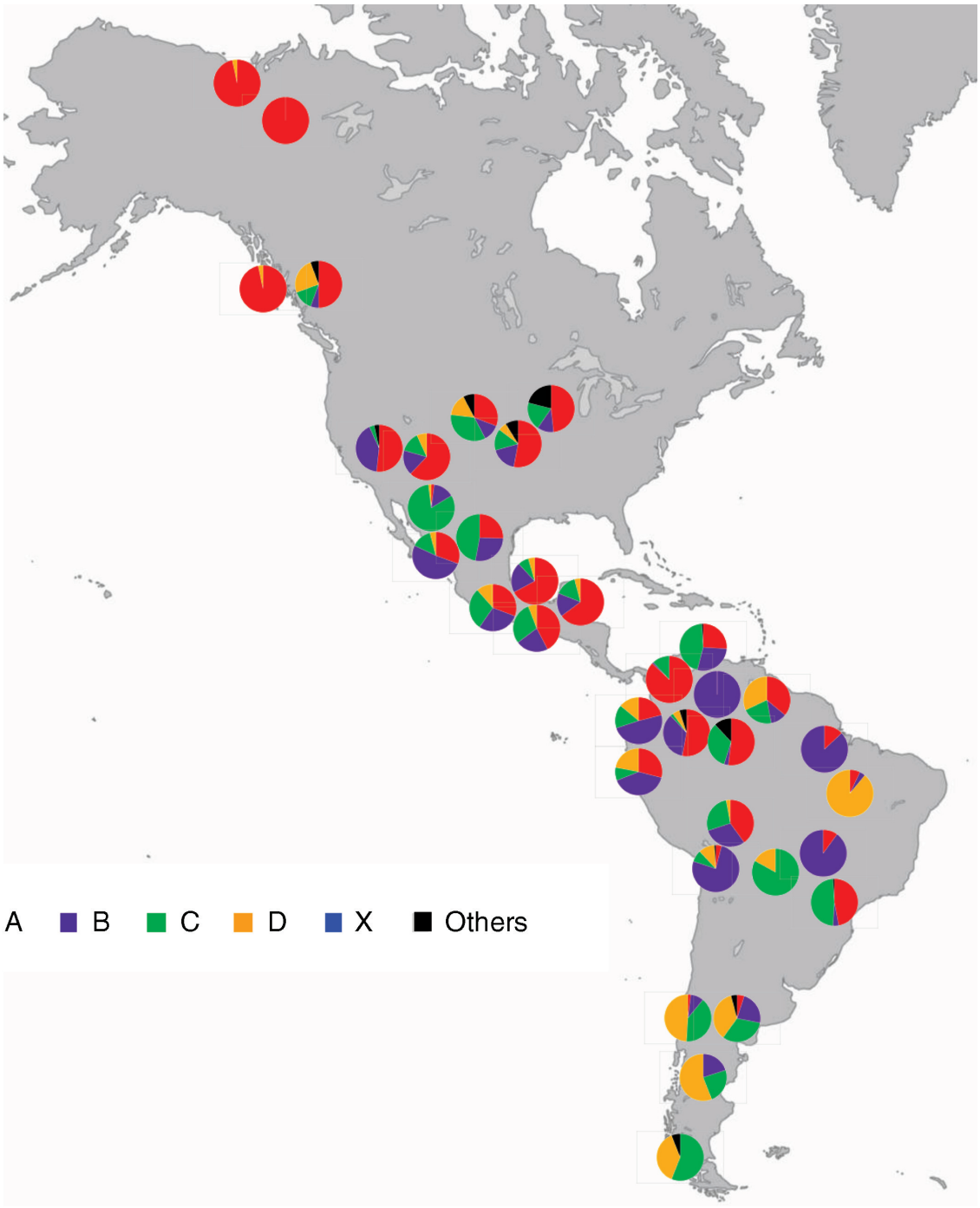


FIGURE 16.13

Map of mtDNA haplogroup frequencies in the New World. Data taken from Torroni, A., et al., "Native American mitochondrial DNA analysis indicates that the Amerind and Nadene populations were founded by two independent migrations," *Genetics* 130:153, 1992; Torroni, A., et al., "Asian affinities and continental radiation of the four founding Native American mtDNAs," *American Journal of Human Genetics* 53:563, 1993; Merriwether, D.A., et al., "Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World," *American Journal of Physical Anthropology* 98:411, 1995; Lorenz, J.G., and Smith, D.G., "Distribution of four founding mtDNA haplogroups among Native North Americans," *American Journal of Physical Anthropology* 101:307, 1996; Keyeux, G., et al., "Possible migration routes into South America deduced from mitochondrial DNA studies in Colombian Amerindian populations," *Human Biology* 74:211, 2002; Kemp, B.M., et al., "Evaluating the farming/language dispersal hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica," *Proceedings of the National Academy of Sciences USA* 107:6759, 2010.

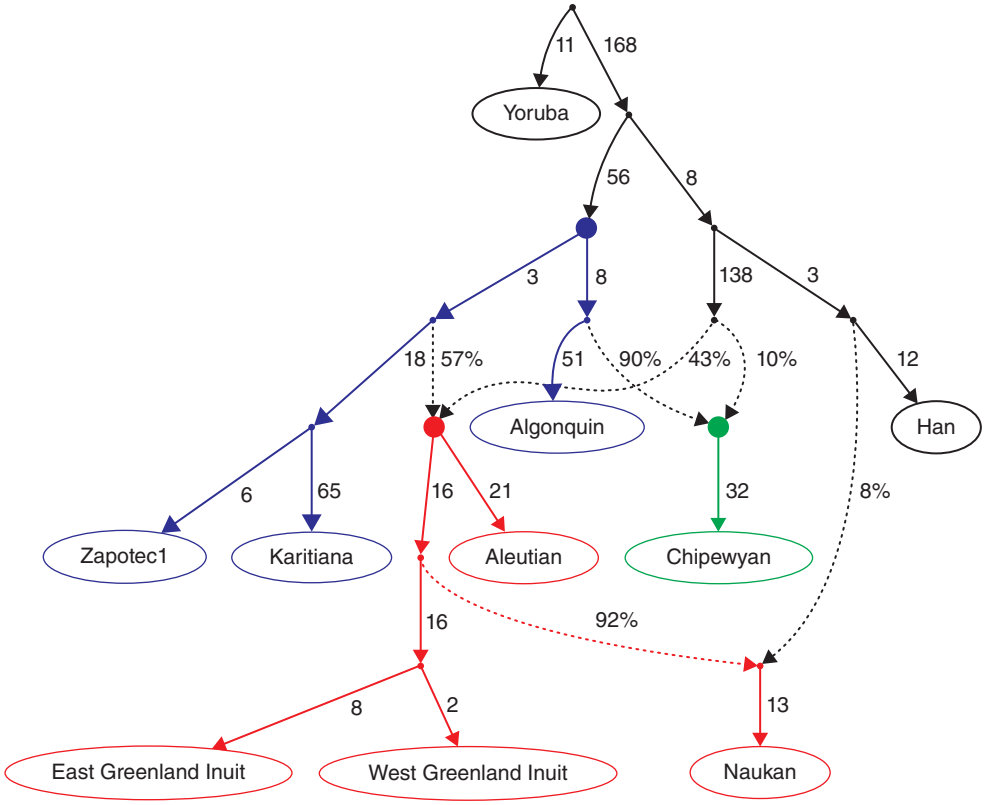


FIGURE 16.14

Admixture graph for the origins of New World populations, based on genome-wide SNP data. Black population names are non-American groups (Han Chinese and Yoruba), blue are the Amerind groups, red are Eskimo-Aleut (including one group from Siberia, the Naukan), and in green is the one Na-Dene group. The numbers on the solid lines are estimated amounts of genetic change (note that branch lengths are not proportional to genetic change); dashed lines indicate migration/admixture events. Reprinted with permission from Reich, D., et al., "Reconstructing Native American population history," *Nature* 488:370, 2012.

identified from ancient remains to date, suggesting that the major founding mtDNA/NRY lineages have indeed been identified in contemporary Native American populations. This latter is an important point for two reasons: first, it suggests that any additional migrations to the New World have had a minimal genetic impact; second, it suggests that the Native American population crash that followed European contact has not dramatically altered patterns of genetic variation in Native American groups. Regarding the second point, it is not known for sure what the precontact population size of the New World was, but estimates for North America range from 4 to 8 million. By the beginning of the twentieth century, there were less than 100,000 Native Americans in North America, and arguably without modern medicine, Native Americans would have gone extinct. This is because while warfare was responsible for some of the loss, by far the major factors in the Native American population decrease were the diseases brought by the Europeans—infectious diseases have always had a major impact on human populations, and resistance to such diseases has been an important source of genetic adaptations in humans (recall the example of sickle-cell anemia and malaria from Chapter 5). It has therefore been a concern that the decreased genetic variation in contemporary Native Americans (relative to most other populations) may reflect the postcontact population crash rather than a small founding population size. Studies of ancient DNA from precontact remains can directly address this question, and to date the studies that have been done have found that the major haplogroups are all represented in current populations (Raff et al. 2011), although there may be some slight decrease in haplotype diversity within each haplogroup (in other words, the postcontact bottleneck may have pruned some twigs, but not any of the major branches, of the tree of Native American mtDNA diversity).

So, to summarize, the genetic evidence seems to support a single major migration to the New World around 15,000 years ago, with possibly additional

migrations that contributed to the ancestry of Eskimo-Aleut and Na-Dene groups. The current favored model posits an earlier separation and isolation of the ancestors of New World groups from Asia, possibly in Beringia (the "Beringian incubation" model), for perhaps as much as 10,000 years before the migration(s) from Beringia to the New World.

INTO EVEN MORE REMOTE LANDS: THE COLONIZATION OF POLYNESIA

If you've never had the good fortune to actually travel through the South Pacific, then you probably don't appreciate the vast distances involved. So, go find a globe and then find the following islands: New Zealand, Hawaii, and Easter Island (or, if you are too lazy to find a globe, see Figure 16.15—just keep in mind that the flat, two-dimensional representation of a map does not fully capture the distances involved). The triangle connecting these three islands forms the boundaries (more or less) of Polynesia. Compare the size of this triangle to North America or Eurasia and you'll start to appreciate just how far people had to travel in order to reach the various islands that make up Polynesia. And, when you then realize that Polynesians were routinely traveling back and forth across these vast distances hundreds of years before European explorers (with their supposedly more advanced sailing technology) dared to venture much beyond the sight of land, well, the colonization of Polynesia was truly a remarkable achievement.

When, from where, and how people got to these far-flung islands has long occupied the attention of many researchers from all sorts of different fields, including archaeology, cultural and physical anthropology, linguistics, and—yes—genetics. And in case you were wondering, the amount of attention that these questions surrounding the colonization of Polynesia have received by academic researchers probably does reflect, at least in part, the fact that tropical islands are a very nice place to do fieldwork (especially if you are in the midst of a dreary German winter, as I am as I write this!). But the good news for geneticists is that all of the information and insights coming from different disciplines has provided a wealth of ideas and hypotheses concerning the colonization of Polynesia and the biological relationships of various populations that can be tested against the genetic evidence.

Let's begin with some terminology and definitions. Historically, the South Pacific has been divided into regions (Figure 16.16), defined partly by geography and partly by culture, called **Melanesia** ("black islands"), **Micronesia** ("small islands"), and **Polynesia** ("many islands"). These names were bestowed by the French explorer Jules Dumont d'Urville (for an English translation of his original essay in French, see Ollivier et al. 2003), based on his observations from

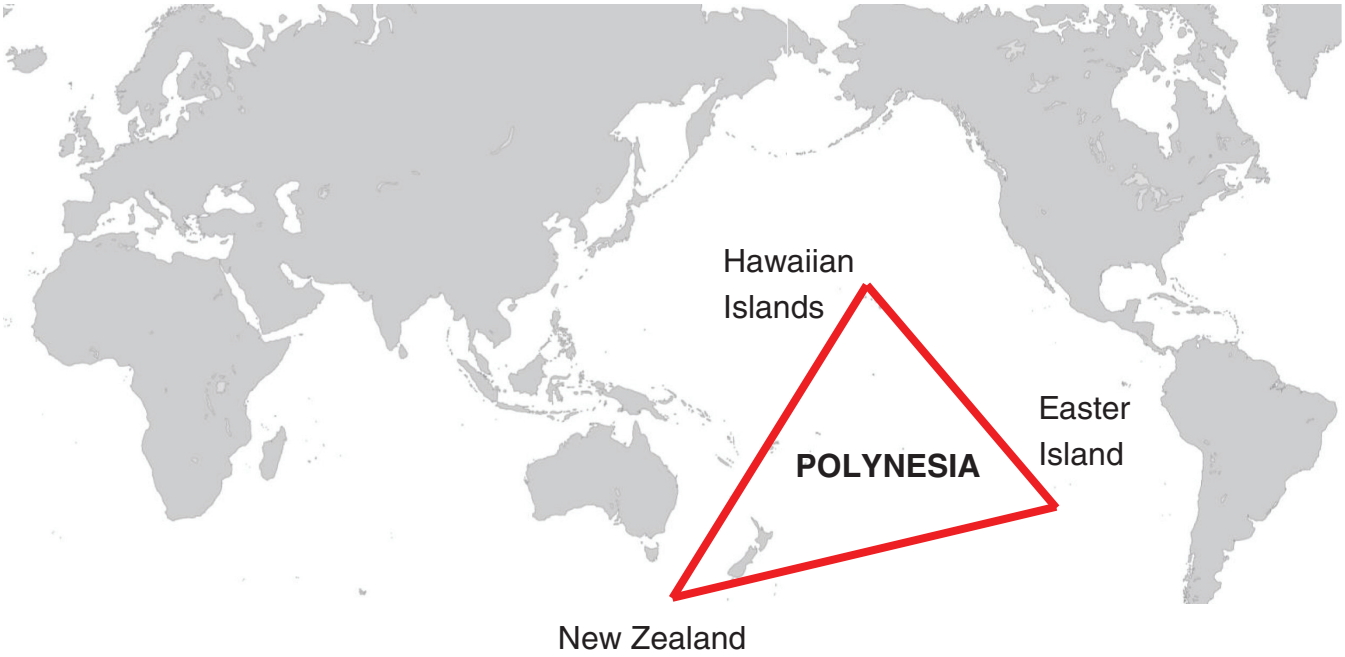


FIGURE 16.15

Map of the world, with the Polynesian triangle (with vertices at New Zealand, Easter Island, and the Hawaiian Islands) indicated. Compare the size of the Polynesian triangle to Europe, Africa, or North America, and you begin to appreciate the vast distances that Polynesians traveled during their colonizing voyages of the Pacific.

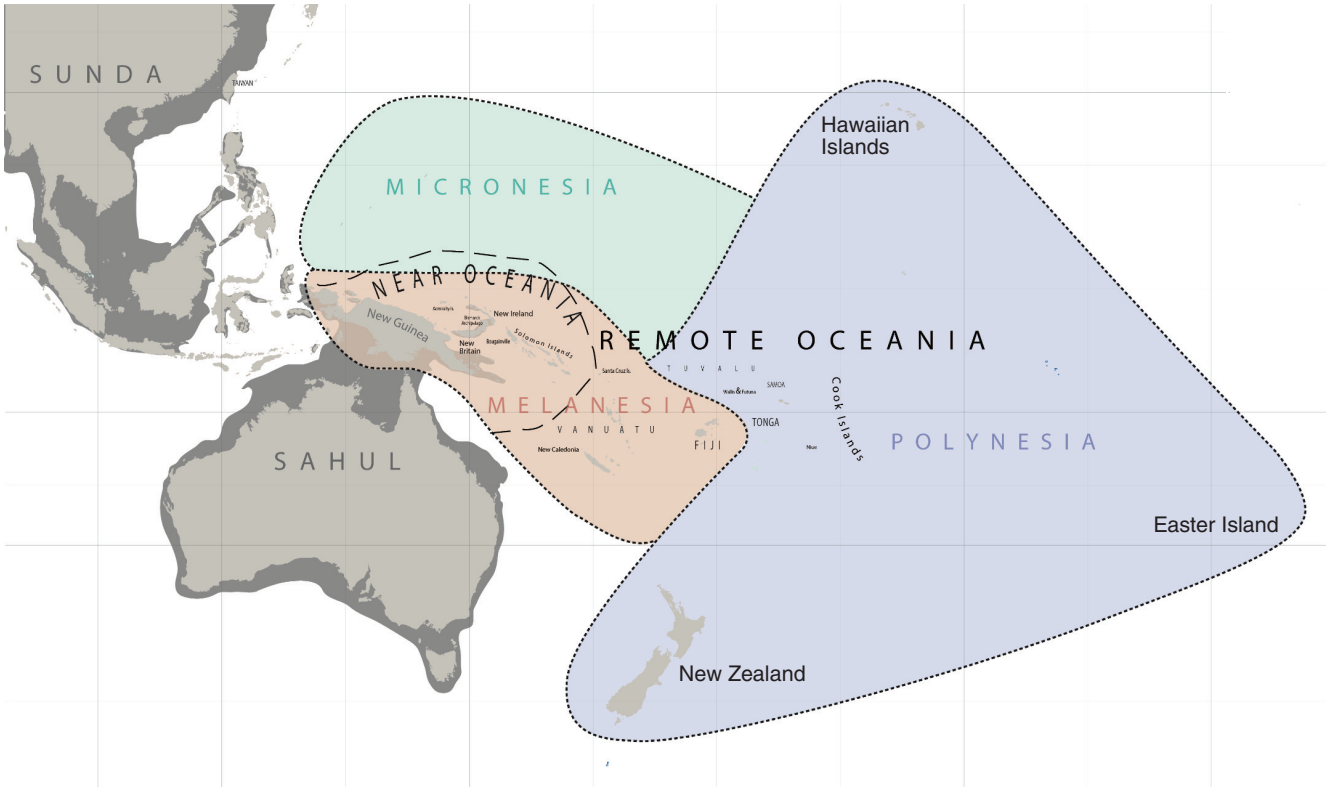


FIGURE 16.16

Cultural areas of the Pacific (Melanesia, Micronesia, and Polynesia); the dashed line shows the boundary between Near and Remote Oceania. The shaded areas around the past continental landmasses of Sunda and Sahul show the extent of land when ocean levels were lower. Provided by and used with the permission of Ana Duggan.

two voyages to the South Pacific in the 1820s. Incidentally, d'Urville had, on an earlier voyage to the eastern Mediterranean, come across a recently discovered statue that he tried to convince the captain of his ship to purchase, but the captain did not share his enthusiasm for antiquities and declined on the grounds that the statue was too bulky to transport. Undaunted, d'Urville persuaded the French ambassador to obtain the statue, and the ambassador's representative managed to seize the statue literally moments before it was to be transported to Constantinople, and brought it to France. It is now considered a national treasure of France, and if you go to the Louvre in Paris, you can see it—it is known as the Venus de Milo. Anyway, Melanesia consists of New Guinea and nearby islands, and the people of Melanesia are characterized by darker skin pigmentation (hence the name) and different cultural traits than Polynesians, while Micronesians show a mixture of Melanesian and Polynesian traits, with some affinities to the nearby Philippines as well.

However, this distinction between Melanesia, Polynesia, and Micronesia is not so satisfactory. For example, in some respects, Fiji has more in common with Melanesia than with Polynesia (a possible explanation for why this might be the case will be forthcoming later). Moreover, it is becoming increasingly apparent that the major distinction between regions in the South Pacific has to do with the time of colonization. Humans entered Sahul (the combined Australia/New Guinea land mass) some 40,000–50,000 years ago and managed to get as far as the main chain of the Solomon Islands by at least 20,000 years ago or so. With the lowered sea levels, all of the island crossings that had to be made to get that far were "intervisible," that is, people would have seen signs of land ahead of them before they lost sight of land behind them. Such crossings presumably did not require sophisticated boating technology; simple dugout canoes (of the sort that one can still see in the islands around New Guinea today) would have sufficed. However, to get beyond Makira, the easternmost of the main Solomon Islands, to the next major island, Santa Cruz, would have required crossing around 400 km of open ocean. Santa Cruz and all islands further to the east (including all of Polynesia) were only colonized much later, beginning around 3000 years ago, and evidently required more sophisticated boating technology (outrigger canoes) and/or navigation abilities. Thus, the major distinction to be made in the South Pacific, in terms of when people reached various islands, is

the distinction between **Near Oceania** and **Remote Oceania** (Figure 16.16): Near Oceania includes Australia, New Guinea, and all islands eastward as far as Makira in the Solomon Islands, while Remote Oceania starts with Santa Cruz (politically also part of the Solomon Islands) and includes all islands further to the east, including all of Polynesia. Although the less precise terms Melanesia, Micronesia, and Polynesia are still frequently used, here we will use the more precise terms, namely, Near and Remote Oceania, unless we want to refer specifically to the islands of the Polynesian triangle.

Over the years, many ideas have been proposed concerning the initial colonization of Remote Oceania and the origin of Polynesians. Most notably, the anthropologist Thor Heyerdahl was convinced that ancient Peruvians, sailing in balsa wood rafts, were the first to reach Polynesia (notwithstanding the fact that Polynesians have never used such rafts). And to demonstrate that this was indeed feasible, in 1947 Heyerdahl went to the remarkable extent of constructing a balsa wood raft, dubbed the Kon-Tiki (named after the Inca Sun god) and setting sail from Peru, ultimately making landfall (well, actually, smashing into a reef) on the island of Rairora in the Tuamotu archipelago. The extraordinary voyage took 101 days and Heyerdahl and his crew of five men and the obligatory parrot survived several hardships along the way. But alas, while the voyage of the Kon-Tiki showed that it was indeed possible for Polynesians to have originated from South America, as discussed below there is absolutely no evidence to support a South American origin for Polynesians. Still, you have to admire Heyerdahl (or wonder about his sanity)—not many scientists would have the courage to go to such lengths and risk their life just to try to prove the feasibility of a hypothesis of purely academic interest.

Before getting to the genetic evidence, let's briefly review the archaeological and linguistic evidence concerning the colonization of Remote Oceania. Figure 16.17 shows the estimated initial colonization times for and migration routes to various islands, based on radiocarbon dates of suitable archaeological materials (bones, shells, plant materials, etc.). These dates show that progressively more recent dates occur as one moves from west to east through Remote Oceania, suggesting that people were moving from west to east. Moreover, there is a characteristic type of pottery found from New Guinea eastward called Lapita, named after the site where it was first found in New Caledonia. This pottery has a very distinctive style, consisting of intricate patterns formed by stamping images on the clay while the pottery is still wet (Figure 16.18); this style of pottery is known as **dentate pottery**. While Lapita pottery is known only from Near and Remote Oceania, related pottery traditions are found in Indonesia, and again the dates for sites with Lapita (or Lapita-like) pottery get progressively younger as

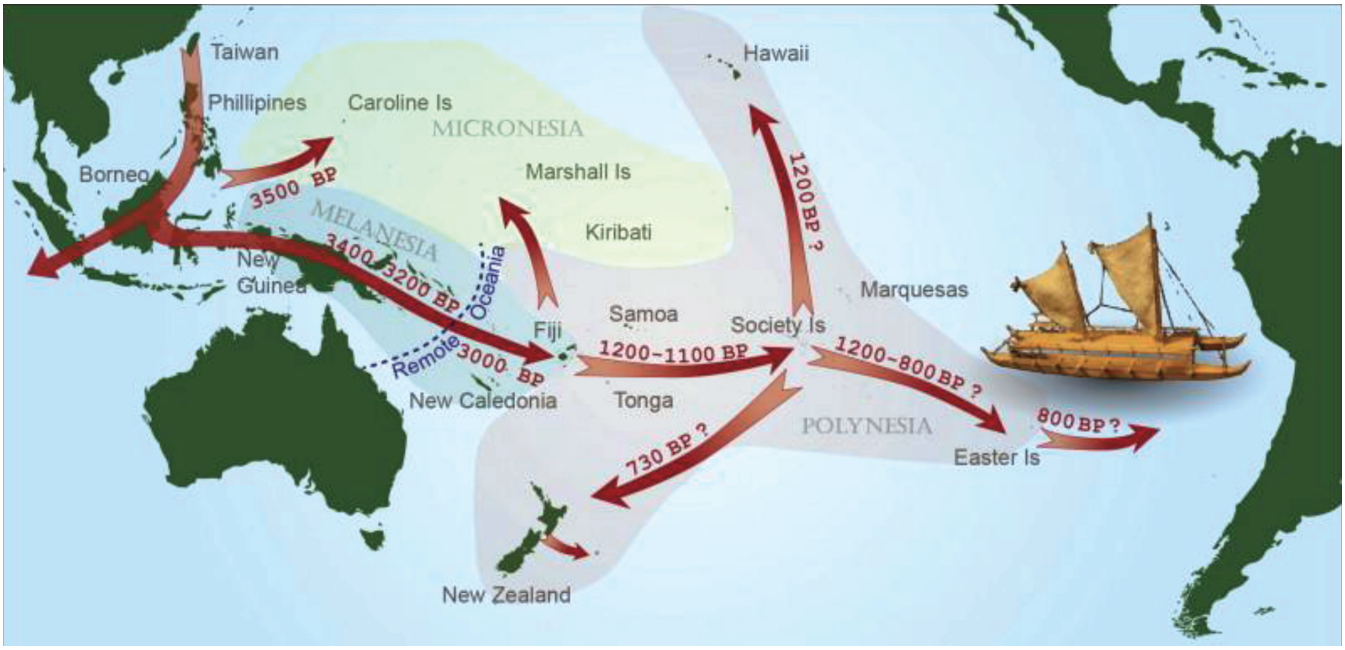


FIGURE 16.17

Inferred time and direction of migrations through the Pacific, based on archaeological data. Reprinted with permission from Matisoo-Smith, E., "Ancient DNA and the human settlement of the Pacific: a review," *Journal of Human Evolution* 79:93, 2015.

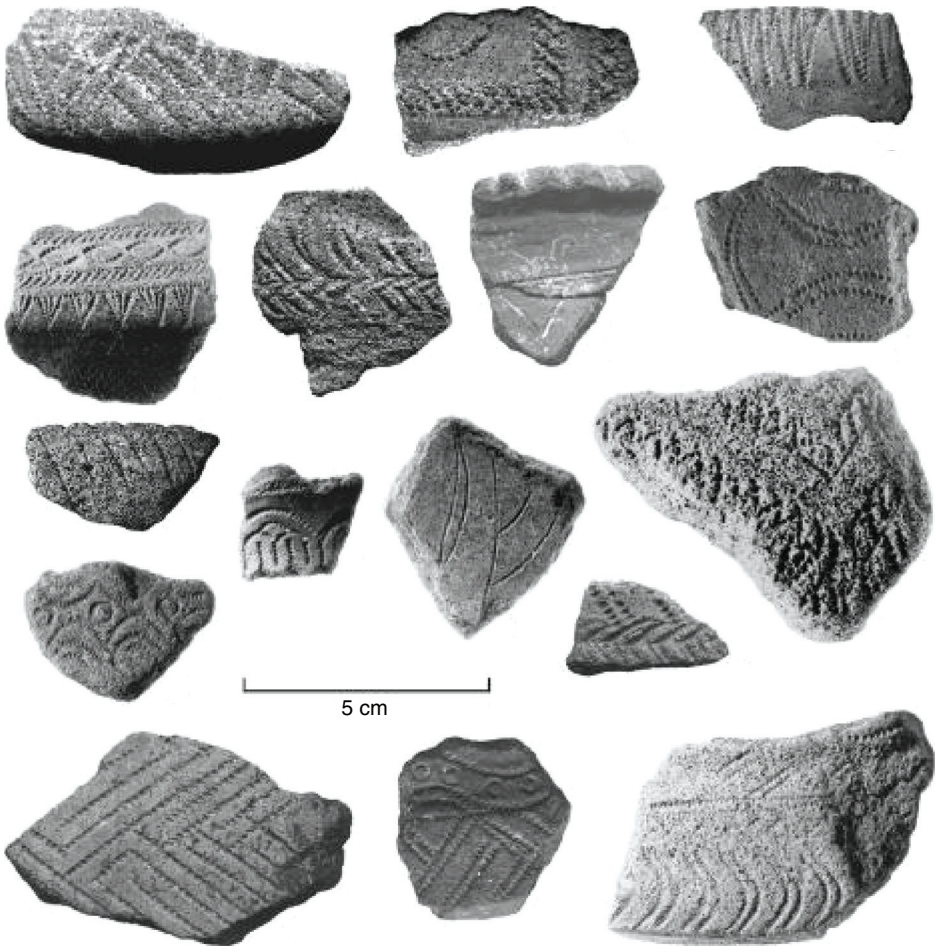


FIGURE 16.18

Lapita pottery sherds from Tonga. Modified with permission from Burley, D.V., and Dickinson, W.R., "Among Polynesia's first pots," *Journal of Archaeological Science* 37:1020, 2010.

one moves from west to east. So, all of the archaeological evidence suggests a west-to-east spread of people from Near through Remote Oceania.

With respect to the linguistic evidence, there are (at least) three major groups of languages in Oceania: Australian, Papuan, and Austronesian. There are currently about 150 Australian languages (out of an estimated 500–700 languages at the time of European contact) found only in Australia. It is not clear whether these all descend from one single ancestral Australian language (perhaps corresponding to the initial colonization of Australia), or whether more than one language was brought to Australia in the past, with different groups of Australian languages tracing to different ancestral languages. All we do know is that Australia harbors a large number of languages for which no connection with the languages of New Guinea (or anywhere else) can be demonstrated, even though there is ample genetic evidence to indicate that the populations of Australia and New Guinea are related, albeit distantly.

The second major group of languages in Oceania are the Papuan languages, an extremely heterogeneous group of some 800 languages—around 15% of the world's known languages—spoken by just a few million people living on New Guinea and nearby islands. These languages are so different from one another that when I started working on the genetics of New Guinea populations as a graduate student in the early 1980s, the languages of New Guinea were classified as either belonging to the Austronesian family (discussed next) or not, with no implication whatsoever that the non-Austronesian languages were related to one another. Lumping languages into a category simply because they do not belong to another category is hardly a satisfactory state of affairs, and recently a lot more work on non-Austronesian languages has shown that many (but by no means all) of them can be related. Non-Austronesian languages are now called Papuan languages, although keep in mind that there are several groups of (so far) unrelated languages subsumed under this term. Papuan languages are mostly spoken

in New Guinea and nearby offshore islands (such as New Ireland and New Britain), extending as far east as the Solomon Islands, with a few Papuan languages also found in eastern Indonesia. It is generally accepted that Papuan languages are likely to be the oldest languages in New Guinea, as evidenced by their extreme heterogeneity and highest concentration in the remote highlands of New Guinea, and may even trace to the initial colonization of Sahul some 40,000–50,000 years ago (but keep in mind that this is just conjecture, as there is as yet no linguistic evidence that can shed light on such time depth).

In contrast to the situation for the Australian and Papuan languages (i.e., extremely heterogeneous groups of very divergent languages), the remaining group of languages in Oceania, the Austronesian languages, are considered to be definitely related to one another and to form a homogeneous family. Austronesian is one of the most widespread language families in the world (Figure 16.19), extending from Taiwan down through the Philippines, Indonesia, parts of coastal New Guinea and nearby offshore islands (interspersed with Papuan languages), through the Solomons, and then throughout Remote Oceania (i.e., all of the languages of Remote Oceania are Austronesian languages). Austronesians were also the first people to colonize Madagascar, which would have involved traveling over 6000 km—those people really got around!

However, one place they did not penetrate was the highlands of New Guinea, as all highland New Guinea groups speak Papuan languages. This is perhaps not so surprising, because the highlands are really quite remote and hard to get into—in fact, it was generally thought that there were no people at all in the highlands, until two Australian gold prospectors hired a plane in 1930 and flew over the highlands and saw, to their astonishment, numerous villages and fields of crops, revealing an extensive population whose existence was completely unsuspected. Reading first-hand

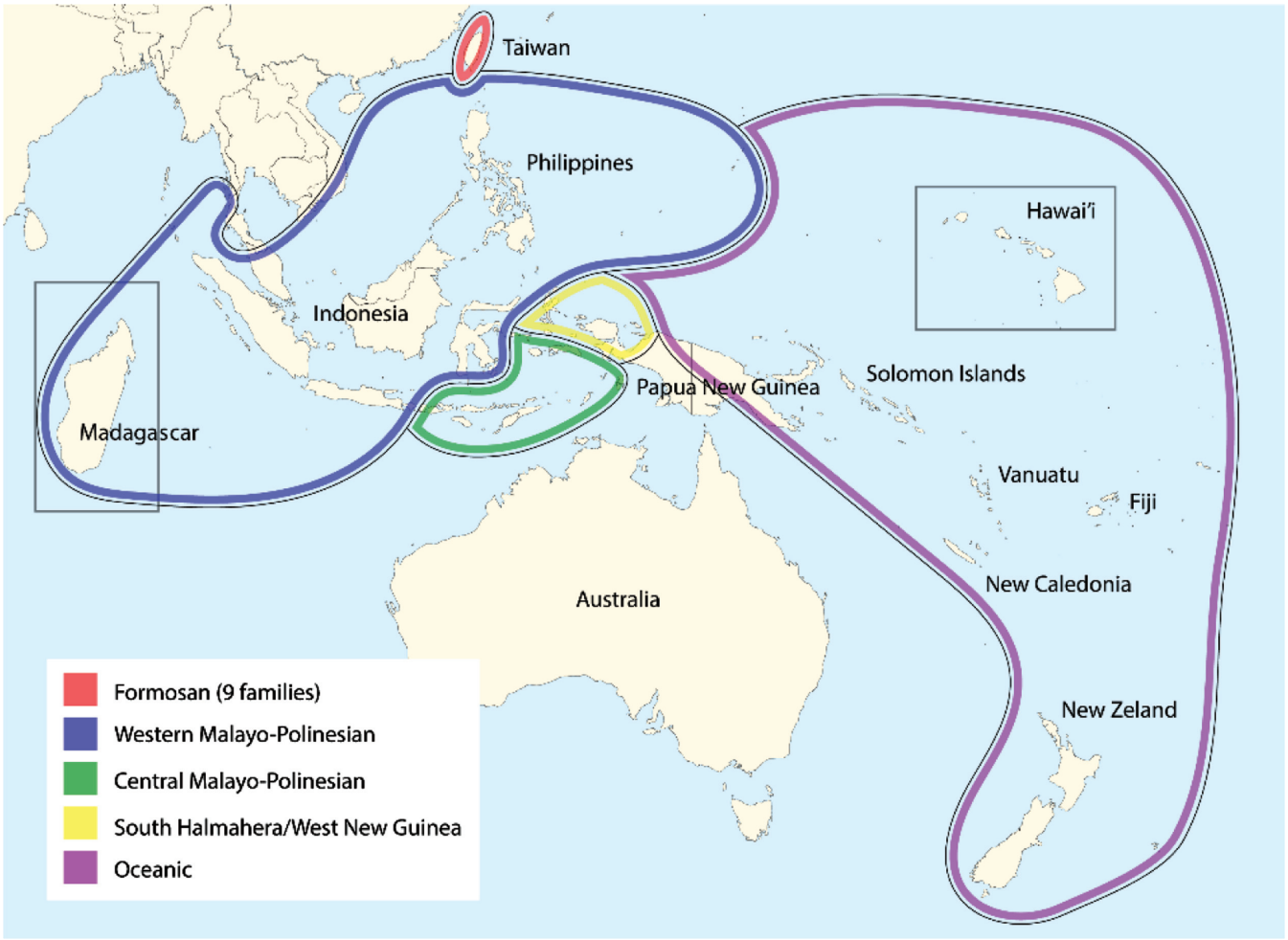


FIGURE 16.19

Map of the distribution of the various subgroups of Austronesian languages. Reprinted with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Austronesian_family.png).

descriptions of the initial contact between highlanders and the European patrols sent to explore the highlands is quite illustrative of the clash of cultures—for example, in one account (Schieffelin and Crittenden 1991) the native women exposed their genitals to the Europeans, as this was powerful magic in their culture and the women expected the strangers to flee in terror. Meanwhile, the Europeans were completely bewildered by this behavior, and wondered whether the women were trying to seduce them! And, hard as it is to imagine in this increasingly globalized and mobile world, there are still groups in highland New Guinea that have never been contacted by outsiders.

Austronesian languages are generally considered to have originated on Taiwan, as linguists have identified 10 different branches of Austronesian languages on Taiwan, whereas all non-Taiwan Austronesian languages (some 1000–1200 languages) are derived from just one of these Taiwan branches. Moreover, by applying Bayesian phylogenetic methods (as described in Chapter 11) to an extensive data set of word lists of Austronesian languages (compiled by linguists), the biologist Russell Gray and colleagues produced a tree of Austronesian languages that shows a remarkable congruence with geography (Gray et al. 2009). As shown in Figure 16.20, there are successive splits in the Austronesian language phylogeny as one moves from Taiwan south through the Philippines and Indonesia and then eastward through Near and Remote Oceania. Gray and colleagues have also dated these splits by assuming a "molecular clock" for languages, calibrated with various linguistic and archaeological time points. The resulting dates suggest that the spread of Austronesian languages began out of Taiwan some 5200 thousand years ago, reached New Guinea by about 3500 years ago, and Remote Oceania by about 1500 years ago. To be sure, the extent to which the phylogenetic and dating methods from molecular evolution can be applied to linguistic data is vigorously debated—recall that to apply such methods accurately, you need an evolutionary model, and while we have a pretty good idea as to how DNA evolves, the same does not hold for how languages change and evolve over time. Nonetheless, the dates for the spread of Austronesian languages are in good agreement with the archaeological evidence discussed previously for Lapita pottery and the initial colonization of the various islands of Remote Oceania (and, as we shall see later, with genetic evidence as well), suggesting that the phylogenetic/dating approach to investigating language relationships does have some merit.

So, the archaeological and linguistic evidence both indicate that the origins of Remote Oceanians lie to the west (i.e., Southeast Asia) and not to the east (i.e., South America—notwithstanding the voyage of the Kon-Tiki). And based on this evidence, several different models for the colonization of Remote Oceania have been proposed, which can be roughly classified into two broad categories. The first of these considers that beginning about 6000 years ago, there was a migration of people from somewhere in East Asia (most likely, Taiwan) that then spread through the Philippines and Indonesia, reached coastal New Guinea and nearby islands about 3500 thousand years ago, and then spread from Near Oceania to Remote Oceania, reaching the farthest Polynesian islands by about 700–800 years ago. These people brought a complete cultural package that included Austronesian languages, distinctive pottery, rice agriculture (at least through island Southeast Asia), and outrigger boating technology. There are a variety of models that fall into this category, including the "Express Train" model of Jared Diamond (which focusses on the rapid spread to Remote Oceania), the "Out of Taiwan" model of Peter Bellwood (which emphasizes the probable Taiwan source of the Austronesian expansion), and the "Triple I" model of Roger Green (which stands for Intrusion, Integration, and Innovation, and hence emphasizes the roles of migration, admixture, and new inventions). While the differences among the details of these various models are not trivial, the models all share the common feature of supposing a definite migration of people from Taiwan/Asia that ultimately reached Near Oceania, and so we will use the generic term "Out of Asia" to refer to these models.

The second broad category of models for the colonization of Remote Oceania denies the importance of any single major wave of migration of people (i.e., the Austronesian expansion) from East Asia to Near and Remote Oceania. According to one such model, after the initial colonization of Near Oceania some 40,000–50,000 years ago, there would have been ongoing biological and cultural interactions between all sorts of different groups in Southeast Asia and Near Oceania, with no single major wave of migration from Asia. Then, beginning around 3000 years ago, people in Near Oceania developed the wherewithal to begin venturing further east, eventually colonizing all of Remote Oceania. This view is most prominently (but by no means exclusively) associated with the anthropologist John Terrell and is sometimes known as the "Entangled Bank" model (after Darwin's use of the term in *On the Origin of Species* to describes the rich diversity of life, the "endless forms most beautiful and most wonderful"). For convenience, we will refer to this category of models as "Out of Melanesia" to emphasize the lack of a single major expansion from Asia that simultaneously brought Austronesian people, languages, and cultural practices to and through Near Oceania before colonizing Remote Oceania.

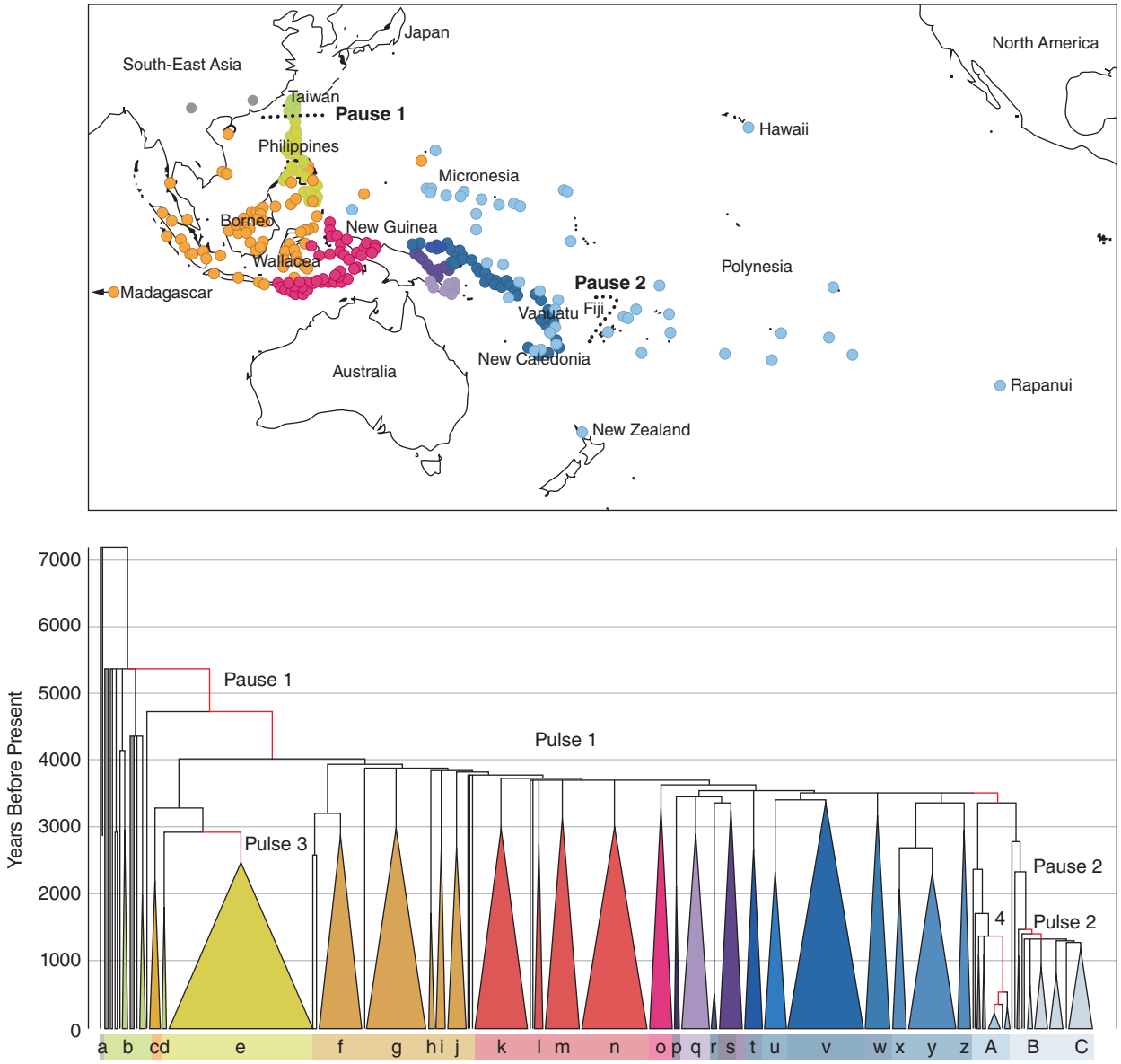


FIGURE 16.20

Phylogenetic analysis of Austronesian languages. Top, map showing locations of languages studied, colored according to which clade in the tree they belong to. Bottom, phylogenetic tree of Austronesian languages, based on lexical data. The tree and map together show a spread of Austronesian languages from Taiwan, beginning about 5000 years ago. From Gray, R., et al., "Language phylogenies reveal expansion pulses and pauses in Pacific settlement," *Science* 323:479, 2009.

We are (at last!) ready to consider the genetic evidence for the colonization of Remote Oceania. MtDNA haplogroups in Near Oceania can be broadly classified (Kayser et al. 2006) into two groups (Figure 16.21). One group consists of haplogroups that are widespread throughout East and Southeast Asia but in Near Oceania are found only in coastal New Guinea and offshore islands—these mtDNA haplogroups are completely absent from highland New Guinea. This distribution thus roughly mirrors the distribution of Austronesian languages in Near Oceania (although these haplogroups are found in both Austronesian and Papuan-speaking groups in coastal and island New Guinea). Moreover, the diversity associated with these haplogroups is higher in Asia than in Near Oceania, suggesting that they most likely originated in Asia (recall from Chapter 12 that the place associated with the highest diversity for an mtDNA or NRY haplogroup is the most likely origin for that haplogroup). We will thus refer to these as "Asian" haplogroups.

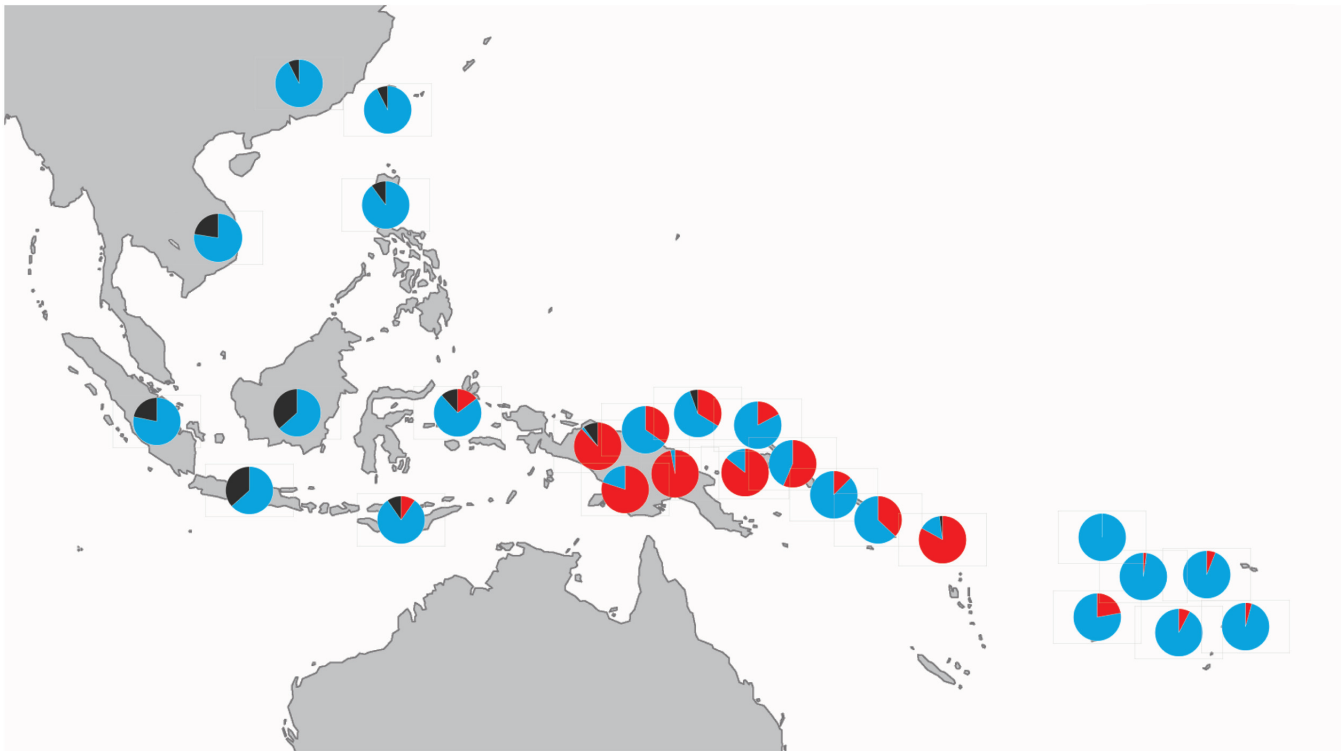


FIGURE 16.21

mtDNA haplogroups in Asia and Near and Remote Oceania, based on HV1 sequences. Red = Near Oceania, Blue = Asia, and Dark gray = Other/unknown. Note that mtDNA haplogroups in Remote Oceania are primarily of Asian origin. Data from Delfin, F., et al., "Bridging Near and Remote Oceania: mtDNA and NRY Variation in the Solomon Islands," *Molecular Biology and Evolution*, 29:545, 2012.

The second group of mtDNA haplogroups (Figure 16.21) are widespread in Near Oceania, occurring both in highland New Guinea groups (where all of the mtDNA haplogroups belong to this class) as well as in coastal New Guinea and nearby islands. However, with the exception of some eastern Indonesian groups, these haplogroups are absent from Southeast and East Asia. The distribution of mtDNA haplogroups belonging to this class thus mirrors the distribution of Papuan languages; moreover, the diversity associated with these haplogroups is highest in New Guinea and lower in eastern Indonesia, suggesting that they originated in New Guinea. We will, therefore, refer to these as "Melanesian" haplogroups.

And what do we find in Remote Oceania? As shown in Figure 16.21, the vast majority (95–100%) of the mtDNA haplogroups in populations from Remote Oceania are "Asian" haplogroups. Moreover, there is extensive sharing of mtDNA types from Remote Oceania, Near Oceania, and East and Southeast Asia (Figure 16.22), indicating a rapid spread of people across the extensive geographic region. The starlike shape of the network of mtDNA types (Figure 16.22) is indicative of a strong population expansion associated with the spread of these mtDNA types. And complete mtDNA genome sequences indicate that the Polynesian mtDNA sequences from the most prevalent mtDNA haplogroup are nested in a clade of the mtDNA phylogeny in which the most divergent lineages are from Taiwan (Figure 16.23), suggesting a Taiwan origin for this clade. In sum, the mtDNA evidence fits very nicely with the "Out of Asia" model for the colonization of Remote Oceania.

To be sure, some aspects of the mtDNA evidence do not fit so neatly with a single "Out of Asia" migration that corresponds to the Austronesian expansion. For example, it has been claimed that the major mtDNA haplogroup in Remote Oceania, B4a1a1a (yes, that is really what it is called—mtDNA haplogroup nomenclature leaves a lot to be desired!), has an estimated age of about 7000 years and the most diversity associated with it is in the Bismarck Archipelago of New Guinea (Soares et al. 2011). Since this predates the arrival of Austronesians in Near Oceania, the authors of this study suggested that this haplogroup was brought to Near Oceania by a pre-Austronesian migration from Asia to New Guinea and nearby islands and expanded only into Remote Oceania later (possibly with the Austronesians). However, while not denying the possibility of pre-Austronesian migrations to Near Oceania,

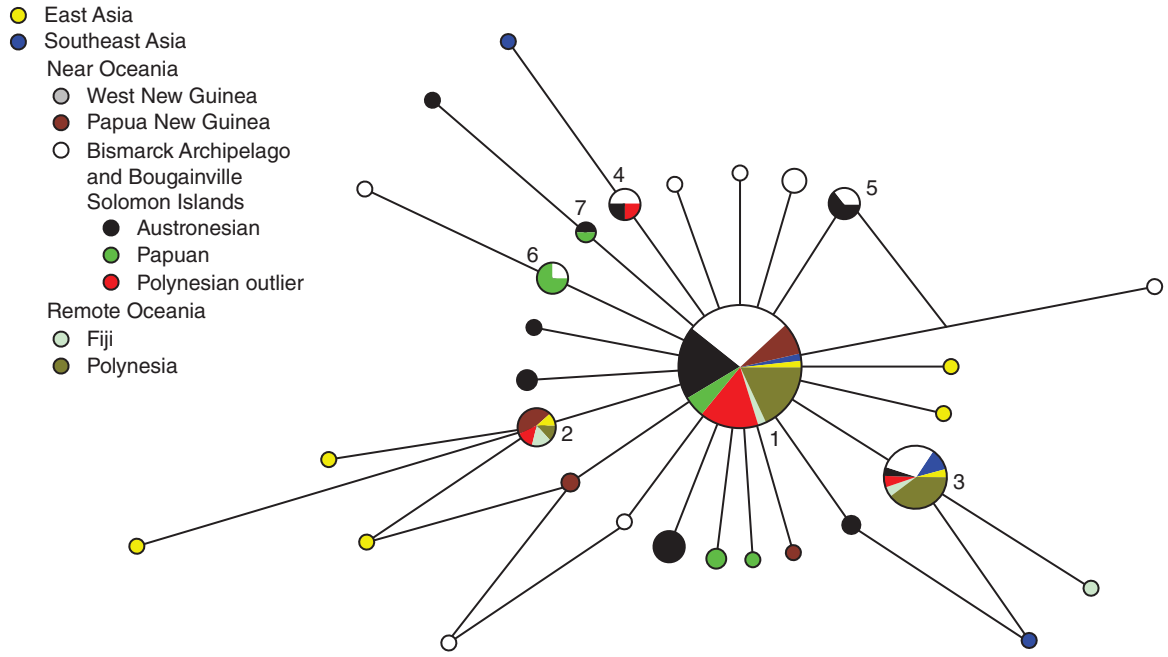


FIGURE 16.22

Network of mtDNA haplogroup B4a1a1 based on HV1 sequences, showing extensive sharing of identical mtDNA HV1 sequences all the way from East Asia to Polynesia. Reprinted with permission from Delfin, F., et al., "Bridging Near and Remote Oceania: mtDNA and NRY variation in the Solomon Islands," *Molecular Biology and Evolution*, 29:545, 2012.

there is enough uncertainty surrounding the dating of the age of the B4a1a1a haplogroup that it's presence in Near Oceania could still reflect the Austronesian expansion (Duggan and Stoneking 2013).

At any rate, the maternal history strongly indicates an Asian source that is predominantly if not exclusively associated with the Austronesian expansion for the colonization of Remote Oceania—what about the paternal history? Analyses of NRY haplogroups in Near Oceania (Figure 16.24) shows that, as with the mtDNA haplogroups, they can be broadly classified into two groups (Kayser et al. 2006), one of probable Asian origin (e.g., NRY haplogroups that are widespread in East and Southeast Asia but restricted in Near Oceania to coastal regions of New Guinea and nearby islands and absent from the highlands), the other of probable Melanesian origin (e.g., widespread in both highland and coastal regions of New Guinea but restricted in Southeast Asia to eastern Indonesia and absent elsewhere in East and Southeast Asia). And what about Remote Oceania? Remarkably, as shown in Figure 16.24, most of the NRY haplogroups in Remote Oceania are of Melanesian origin—overall, about 66% of the NRY haplogroups are of Melanesian origin, which would seem to support the "Out of Melanesia" model for the colonization of Remote Oceania.

So what gives—how can we explain the fact that 94% of the mtDNAs in Remote Oceania are of Asian origin, while 66% of the Y chromosomes are of Melanesian origin? This is certainly one of the largest discrepancies between mtDNA and NRY origins found in human populations. Well, one possibility is that females left Asia and males left Near Oceania, and they then met in Remote Oceania, but somehow that doesn't seem like a very good model. Instead, Manfred Kayser and I proposed the following model (Kayser et al. 2000) to explain the apparently contradictory mtDNA and NRY results: there was an Austronesian expansion out of Taiwan that spread through the Philippines and Indonesia, and then along the coast of New Guinea and offshore islands, but these people didn't simply speed through Near Oceania on their way to Remote Oceania. Instead, they spent some time in New Guinea and admixed with the Papuan groups, and as a consequence of this admixture they picked up "Melanesian" Y chromosomes (and, to a much lesser extent, Melanesian mtDNAs), while also leaving behind their "Asian" mtDNAs (and, to a much lesser extent, their Asian Y chromosomes) in coastal/island Near Oceania. After this pause and admixture, which may have been for only a few generations, the migration eastward continued, leading ultimately to the colonization of the most remote islands of Polynesia. To emphasize the distinction between this expansion and admixture model and the "express train" model favored by Jared Diamond and others, we called this

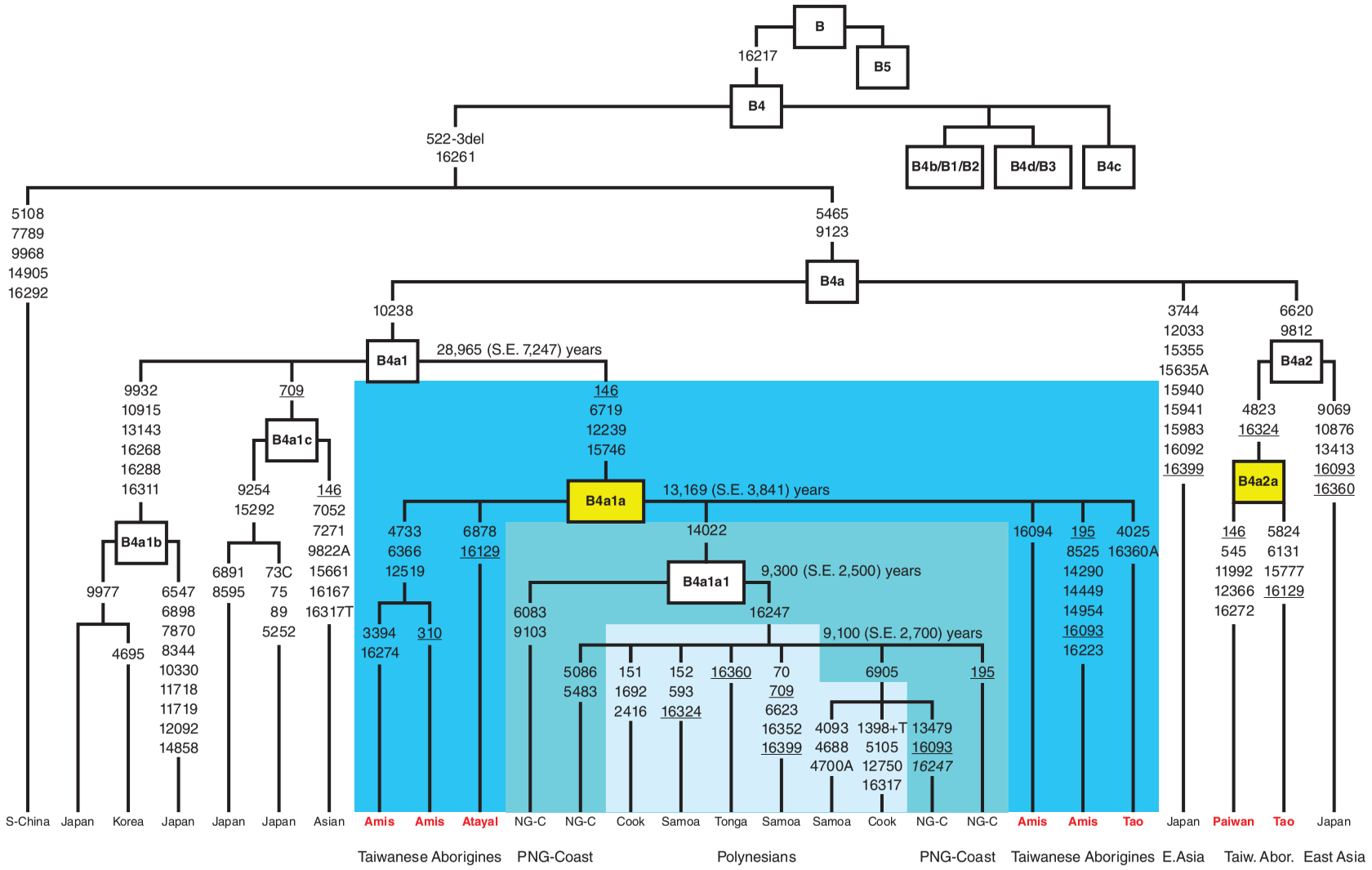


FIGURE 16.23

Phylogeny based on complete mtDNA genome sequences, showing that New Guinean and Polynesian B4a1a1 sequences are nested in a clade of Taiwan mtDNA sequences, thus suggesting a Taiwan origin. Reprinted with permission from Trejaut, J., et al., "Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations," *PLoS Biology* 3:e247, 2005.

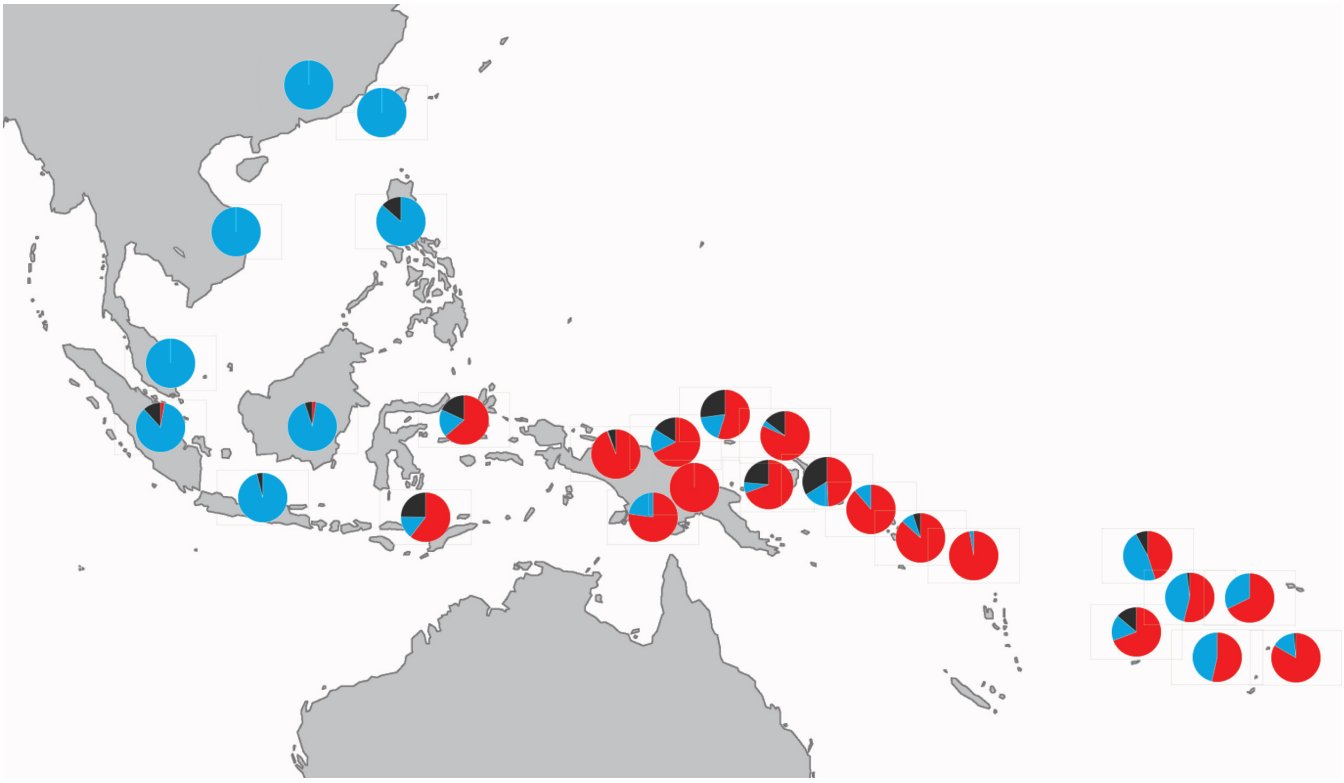


FIGURE 16.24

NRY haplogroups in Asia and Near and Remote Oceania. Red = Near Oceania origin, Blue = Asia origin, and Dark gray = Other/unknown origin. Note that in contrast to the mtDNA results (Figure 16.21), the majority of NRY haplogroups in Remote Oceania are of Near Oceanian origin. Data from Delfin, F., et al., "Bridging Near and Remote Oceania: mtDNA and NRY variation in the Solomon Islands," *Molecular Biology and Evolution*, 29:545, 2012.

the "slow boat" model for the colonization of Remote Oceania—although unfortunately this same name was used shortly thereafter to refer to a different model that emphasized Indonesia as a probable source for the expansion to Remote Oceania (Oppenheimer and Richards 2001).

A key feature of our slow boat model is the sex-biased nature of the admixture that occurred in the ancestry of Polynesians, involving primarily Melanesian males and Austronesian females. At first glance, this might seem to contradict the general expected pattern of males from the incoming, more technologically advanced migrating group having access to females of the resident group but not vice versa (i.e., hypergyny, as mentioned in Chapter 15 in the context of the mtDNA evidence for interbreeding between Neandertals and modern humans). However, what we observe in the patterns of mtDNA and NRY variation in Remote Oceania would be expected if the migrating Austronesians were **matrilocal**, meaning that they preferred to incorporate males, rather than females, from other groups. And indeed, there is some evidence to suggest that ancestral Austronesians may have had a matrilocal social structure: for example, the anthropologist Fiona Jordan concluded, from a reconstruction of ancestral residence patterns, that it is likely that the incoming Austronesians did have a matrilocal social structure (Jordan et al. 2009). We will revisit the issue of how social structure can impact patterns of genetic variation in Chapter 19; for now, regardless of the underlying reason, sex-biased admixture seems to have played a key role in shaping mtDNA versus NRY variation in Remote Oceania.

While comparisons of mtDNA and NRY variation are quite useful for investigating sex-biased processes, genome-wide data are necessary for providing further insights into demographic history. The first such study, carried out by the anthropologist Jonathan Friedlaender and colleagues in 2008 (Friedlaender et al. 2008), made use of the services of the Marshfield Clinic (mentioned back in Chapter 7) to analyze several hundred autosomal STR markers in populations from Near Oceania (New Guinea, New Britain, New Ireland, and Bougainville), Remote Oceania (Polynesia and Micronesia), Taiwan, and East Asia (and one European group). The results indicated that, overall, Polynesians are indeed of mixed Asian and Near Oceanian ancestry. Moreover, there is a signal of ancestry

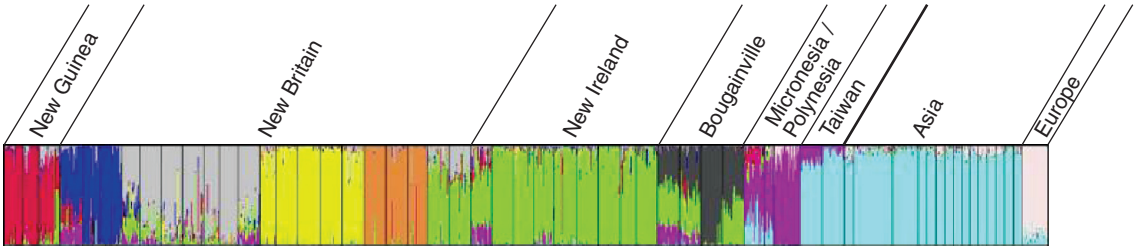


FIGURE 16.25

STRUCTURE analysis based on genome-wide autosomal STR loci for populations from Europe, Asia, Taiwan, and Near and Remote Oceania. See text for further details. Reprinted with permission from Friedlaender, J., et al., "The genetic structure of Pacific Islanders," *PLoS Genetics* 4:e19, 2008.

in the groups from Remote Oceania coming specifically from Taiwan—take a look at Figure 16.25, which shows the results of a STRUCTURE analysis for $k = 10$ (i.e., assuming that 10 different ancestry components can account for the observed genetic structure). Recall from Chapter 11 that in a plot like that shown in Figure 16.25, each individual line represents the ancestry of each individual, with different colors representing the 10 different assumed ancestral components, and that the ancestry for each individual is assigned purely from the genetic data, without utilizing any information concerning the population that the individual belongs to. This plot is quite colorful and quite complicated, with lots of genetic structure evident—for example, there are five different ancestry components that predominate in just the different groups from New Britain. Of interest to us is the purple ancestry component, which is at highest frequency in Micronesia and Polynesia, and also shows up in appreciable frequencies in Taiwan and a few Near Oceania groups, but which is essentially absent elsewhere. A reasonable explanation for this pattern is that people bearing the purple ancestry component spread from Taiwan, left some of their genes in Near Oceania, and then continued on to Remote Oceania, where the purple component increased in frequency as a consequence of genetic drift resulting from bottlenecks and founder events during the colonization of Remote Oceania. But keep in mind that this is a purely descriptive analysis—other explanations are possible, and the Taiwan origin hypothesis was not rigorously tested against competing explanations.

Additional insights came from the subsequent study described earlier in this chapter by Wollstein et al. (2010), which to remind you used genome-wide SNP data to address the issue of single versus multiple dispersals of modern humans from Africa (cf. Figure 16.5). This study used the same approach to test several different models for the colonization of Remote Oceania. After determining the best-fitting model, some parameters related to demographic history were inferred, again using summary statistics to choose the parameter values that gave the closest fit between the simulated data and the observed data. The best-fitting model and associated parameter values (Figure 16.26) look rather complicated (and it is!), so let's walk through it. The first split in the tree involves the single out-of-Africa dispersal and dates to about 55,000 years ago, with the next dispersal to New Guinea, as we saw earlier. A second dispersal from the non-African source population then gives rise to other Eurasians (represented by Europeans, Han Chinese, and Borneo), as inferred previously. The divergence time between New Guinea and Eurasia was dated to about 27,000 years ago, which is somewhat more recent than would be expected given the initial colonization of Sahul around 40,000–50,000 years ago and may indicate that the correction for ascertainment bias in the SNP data is not fully adequate to deal with all the vagaries of ascertainment bias.

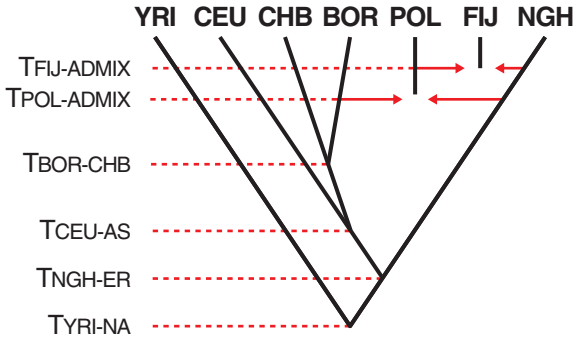


FIGURE 16.26

A model based on genome-wide data supporting the admixed history of Polynesians and Fijians. See text for details. Populations are: YRI, Yoruba; CEU, European; CHB; Han Chinese; BOR, Borneo; POL, Polynesia; FIJ, Fiji; and NGH, New Guinea Highlands. Modified with permission from Wollstein, A., et al., "Demographic history of Oceania inferred from genome-wide data," *Current Biology* 20:1983, 2010.

Unfortunately, the study did not include any aboriginal Taiwan groups, so the Borneo group was used as the closest proxy, which at least makes sense in that if the ancestors of Remote Oceanians did come from Taiwan, they would probably have passed through Borneo. And reassuringly, Borneo does provide a significantly better fit than Han Chinese for the Asian ancestry in Polynesians. So, the best-fitting model depicts Polynesians as having about 87% of their ancestry from Borneo and about 13% from New Guinea. Moreover, the estimated time for the admixture between Borneo and New Guinea is about 3000 years ago, in astonishingly good agreement with archaeological and linguistic evidence for the arrival of Austronesians in New Guinea. And, the effective population sizes are estimated to be about 4000 for Borneo, 2000 for New Guinea, and 1000 for Polynesia, which seem reasonable—certainly we would expect the largest population sizes in Asia and the smallest in Polynesia.

Of perhaps more interest are the results concerning Fiji, because here we learn something new. The best-fitting model indicates that while Fijians and Polynesians have ancestry from both New Guinea and Asia (as described in the previous paragraph), Fijians have additional ancestry from New Guinea that is not found in Polynesians. Moreover, the estimated time for this additional admixture from New Guinea is around 500–1000 years ago. How do we interpret these results? The most likely scenario is as follows: Austronesians arrived in New Guinea about 3000 years ago, mixed with the locals, and carried both Asian and New Guinea ancestry to Fiji. From Fiji, humans spread throughout the rest of Remote Oceania, ultimately colonizing the most remote Polynesian islands. But after the ancestors of Polynesians left Fiji (at least 1500 years ago), there were additional migrations from New Guinea (or elsewhere in Near Oceania, such as the Solomon Islands) to Fiji that contributed additional New Guinea ancestry to Fijians but not to Polynesians. It is likely that this genetic contact occurred over a period of many generations, as the estimated amount of New Guinea ancestry in Fijians varies quite widely (from 22% to 63%), indicating that different individuals have quite different amounts of New Guinea ancestry. By contrast, there is a much narrower interval for the amount of New Guinea ancestry in Polynesians of about 18–28%, which is in line with a single major dispersal of Austronesians to Polynesia. Moreover, this additional genetic contact between Near Oceania and Fiji, that did not spread further than Fiji, is likely to be responsible for the greater similarities observed phenotypically and culturally between Melanesia (i.e., New Guinea and nearby islands) and Fiji (indeed, on this basis Fiji is considered by some to be part of Melanesia, despite being located in Remote Oceania) than between Melanesia and Polynesia.

Taken all together, the genetic, linguistic, and archaeological evidence thus seem to indicate that there was a complete package of genes, languages, and cultural traditions that migrated in the form of Austronesian people from Taiwan (or perhaps from elsewhere in East Asia, although this is less likely) through island Southeast Asia and along the coast of New Guinea and nearby islands, mixing with people along the way, before colonizing the previously uninhabited islands of Remote Oceania. To be sure, while this is the view that I favor, not all researchers agree; some have tried to point to perceived inconsistencies and discontinuities between the genetic, linguistic, and archaeological evidence that, in their view, indicates a much more complicated process behind the colonization of Near and Remote Oceania than the relatively simple picture I just outlined. For instance, we can't know for sure what language was spoken by the people who left behind agricultural implements or Lapita pottery, even though there is a tendency to assume that all makers of Lapita pottery were Austronesian speakers; in other words, just because someone drives a Volvo doesn't mean that he or she is Swedish. But, while driving a Volvo alone does not make you a Swede, if in addition you eat distinctively Swedish food, live in a distinctively Swedish-type society, and have a wide collection of Swedish cultural artifacts, then there is a rather high probability that you are indeed Swedish (Greenhill et al. 2010)—to which I would add that if you also have genes that trace back to Scandinavia, the probability is even higher.

Even though there is compelling evidence that Polynesians are of mixed Asian–Melanesian ancestry, clearly indicating that their origins lie to the west, was there ever any contact between Polynesia and South America? Did people make it from Polynesia to South America (or vice versa)? In fact, there is some evidence from nonhuman sources to suggest such contact. One such source is the sweet potato. The sweet potato was domesticated in the New World, and even though European explorers brought the sweet potato to much of Southeast Asia, there is ample evidence of sweet potato remains in archaeological sites in Polynesia that predate the European explorers (Roullier et al. 2013). And while it is possible that sweet potato tubers rafted from South America to Polynesia on floating vegetation, or via drifting in unmanned boats, the linguistic evidence favors direct human transport of sweet potatoes from South America to Polynesia: the name for sweet potato in Polynesian languages (*kumara*) would appear to be too similar to the name for sweet potato in some South American languages (such as *cumal* or *cumar*) to have arisen by chance—unless the sweet potatoes that drifted on their own carried labels….

Another potential example is the chicken, which is of Southeast Asian origin and was generally thought

to have been introduced to the New World via European explorers. However, a few years ago a study made headlines by claiming that a chicken bone found at an archaeological site in Chile was dated to about 100 years before the arrival of Europeans, and moreover DNA analysis showed that the Chilean chicken bone had an mtDNA type also found in Polynesia (Storey et al. 2007). The conclusion: Polynesians brought chickens to the New World in pre-Columbian times (leading to the inevitable jokes about why did the chicken cross the Pacific Ocean, or Polynesians coming to South America for a picnic and leaving their trash behind). But a follow-up study (Gongora et al. 2008) raised questions about the accuracy of the dating—the chicken bone could easily be younger and overlap with European contact. Moreover, this second study analyzed a much larger sample of contemporary chicken mtDNA sequences and found that the ancient Chilean chicken bone mtDNA sequence is not exclusive to Polynesia but actually is the most common chicken mtDNA sequence in the world and is widely distributed across Europe and Asia (the authors of the study called this the "Colonel Sanders" chicken mtDNA type). Thus, the chicken bone could easily have come from chickens brought to South America by European explorers and not by Polynesians. There have been further studies, replies, and counterreplies by the groups involved in these two studies; at last count no fewer than nine papers have appeared in the prestigious journal "Proceedings of the National Academy of Sciences USA," all concerned with what this one little chicken bone might mean, and while the latest evidence would seem to argue against a pre-Columbian source of this chicken bone (Thomson et al. 2014), we probably haven't seen the last word.

What about the human genetic evidence? Superficially, Polynesians and Native Americans do share some genetic markers, but this reflects the fact that, as we have seen, both groups trace their ancestry back to Asia. For example, mtDNA haplogroup B is found in both Polynesia and the New World, but as different subhaplogroups (B4a and sublineages in Polynesia, B2 and sublineages in the New World). Some specific Native American alleles have been detected in some Polynesians (e.g., Hurles et al. 2003), but the islands that these are found on coincide with those involved in slave trade with Peru during the nineteenth century, which may account for their presence. However, two recent studies have provided convincing genetic evidence of contact between Polynesia and South America, but the crucial interpretation of this contact (as pre- or post-European) rests on aspects of the dating. The first study, based on genome-wide SNPs in various Polynesian and Native American populations, found Native American ancestry in Easter Islanders but no other Polynesian population (Moreno-Mayar et al. 2014). The admixture was dated to 1280–1495 AD by considering the lengths of the blocks of Native American ancestry in the Easter Islanders (see Chapter 12 for more on how this is done), which the authors argued was support for pre-Columbian contact. Fair enough, but as I have tried to stress throughout this book, all such analyses make a number of assumptions, and if any of these are off, even by a little bit, then the admixture date could become more recent and maybe reflect European voyages such as the slave trade. Still, since Polynesians clearly reached Easter Island before the Europeans, it seems reasonable to think they wouldn't have stopped there, but would have continued eastward, and so for sure they would have landed on South America, and for sure they would have been able to make their way back to Easter Island. So, it seems entirely plausible that some Native American ancestry could have ended up in Easter Island. One interesting open question is how far into Polynesia this Native American ancestry extends; the only other Polynesians in this study came from western Polynesia, so it would be useful to know whether Native American ancestry occurs elsewhere in eastern Polynesia.

The second study concerns the rather startling find of two Polynesian skulls among a museum collection of skulls attributed to the Botocudo Indians of Brazil. Initially, these two skulls were found to have Polynesian-like mtDNA sequences (Gonçalves et al. 2013); genomic sequencing confirmed that they had 100% Polynesian ancestry (Malaspinas et al. 2014). The explanation that immediately jumps to mind is a mix-up in the museum collection, but these two skulls are part of a numbered series, all attributed to the Botocudos, all showing the same general preservation, and all other skulls tested showed 100% Native American ancestry, so a museum mix-up is highly unlikely. Plus there is no record of the museum having Polynesian skulls at the same time. Radiocarbon dating of the remains shows that it is unlikely, but barely possible, that these two individuals were alive when Europeans started voyaging to Brazil—even though there is no record of ships going from Polynesia to Brazil during these times. Even if they were somehow transported on a European ship—either as part of the crew, or perhaps as stowaways—there is a remarkable story here as to how they would have jumped ship, made their way inland, and joined the Botocudos. And it is even more remarkable if they indeed made their way on their own from Polynesia to Brazil, either traveling overland or sailing around the tip of South America—surely a Hollywood blockbuster in the making.

One final question about the colonization of Remote Oceania: why did they do it? What on earth possessed people to get into relatively frail boats, not more than 20 meters long, and cross vast distances with no idea as to what they would find, or if they would even

survive? And while some people have suggested that some of the voyaging was accidental, make no mistake about it, they knew what they were doing and what they were up against. There is plenty of evidence to indicate that even the most far-flung islands of Remote Oceania were not colonized just once on a hit-or-miss basis, but instead there was regular travel between the islands. Archaeological evidence indicates well-established trade networks among islands (Kirch and Kahn 2007), and even genetic evidence provides support for this idea, in the form of a Y chromosome marker that seems to have arisen in Polynesia and moreover is widespread across Polynesia, and hence was probably spread by contact between different islands (Cox et al. 2007). So what motivated them to undertake such risky behavior? Genetics would not seem to be the place to look for an answer to this question—except that it has been suggested that one of the characteristics of humans is an innate desire to see what is out there, and that this behavior was perhaps even selected for during human evolution and was responsible for the successful spread of our species around the globe. I have to confess that I am rather skeptical of the idea that humans have a "wanderlust" gene—or at least, I was until I heard a talk a few years ago by Marc Heppener from the European Space Agency, describing what is involved with getting humans to Mars and back again. It turns out that getting humans to Mars is not so difficult, and practically all of the necessary technology already exists. The hard part is getting humans back from Mars alive and in one piece, which would seem to be a formidable obstacle—but Heppener mentioned that whenever he talks about this with astronauts, they invariably say, then go ahead and send us to Mars, we don't care if we come back or not, we just want to go and see what's out there! Such thinking may explain the motivation of the original colonizers of Polynesia (and for sure explains why I'm not an astronaut!).

SOME CONCLUDING REMARKS

In the previous discussions concerning the genetic evidence for the colonization of the New World and the Pacific, I have taken the view that the evidence supports relatively simple models (e.g., a single major colonization of the New World; a major impact of the Austronesian expansion in Near and Remote Oceania, etc.). This reflects both my personal preference for simple versus complex models, plus the fact that science is an inherently reductionistic business: if you cannot reject a simple model based on the evidence at hand, then for sure you cannot reject more complex models. Critics of these simple models for the colonization of the New World and Pacific (and there are many!) would argue that there are many details that don't fit with these simple models that I am sweeping under the rug, and, moreover, humans are rather complex creatures, so we should in fact expect complex rather than simple models for how they migrate. I certainly agree that the models described previously are inherently overly simplistic: humans do not migrate all at once in a single burst but rather in trickles over a period of time, and undoubtedly there were numerous migrations to both the New World and the Pacific that did not leave a trace in the genetic record of current populations. But the real value of the models described previously is not as a representation for what actually happened in the past but rather as a description of the dominant signal(s) present in the gene pools of contemporary populations. Moreover, once we have figured out the dominant genetic signals, we can then start to look for signals of other events. The techniques for doing so are still pretty crude, but they are getting better, and we are starting to get a glimpse of these other events. For example, while the dominant signal in the genomes of native Australians is of an "early southern route" migration (as described previously in this chapter), we recently detected a signal of a migration from India to Australia, dated to about 4000 years ago (Pugach et al. 2013). Intriguingly, microliths (small, finely worked stone tools) and the dingo both show up in the archaeological record for Australia at about this same time, which then raises the possibility that these events may all be related. I expect that in the next few years, we will see many more signals of previously unexpected migrations coming from analyses of genome-wide data.

In addition, we can expect to learn more about the migrations for which we already have evidence. For example, it is well-known that Bantu-speaking agriculturists arrived in southern Africa around 2000 years ago and admixed with the resident Khoisan hunter-gatherers, and this would be the simple version of the story. However, a recent study of genome-wide SNP data from a large number of southern African Khoisan groups shows that this admixture actually happened at different times within the past 2000 years and contributed different amounts of Bantu genetic material to different Khoisan groups (Pickrell et al. 2012). Moreover, there is growing evidence from genetics to support a migration (originally proposed from linguistic and archaeological evidence) of pastoralists from eastern Africa to southern Africa, shortly before the expansion of Bantu speakers reached southern Africa (Breton et al. 2014; Macholdt et al. 2014). In fact, it appears that this migration brought Eurasian ancestry to southern African populations, as the migrating groups from eastern Africa had previously admixed with Eurasians around 3000 years ago (Pickrell et al. 2014)—yet another example of unexpected ancestry

showing up. So, the simple version of the story (Bantu-speaking agriculturists admixed 2000 years ago with Khoisan hunter-gatherers) gets more complex, and we can now try to understand what geographic, social, or other factors may have influenced how different Khoisan groups interacted with the Bantu-speaking agriculturists (and/or with pastoralists from eastern Africa). So, the simplistic (or, as some would say, simple-minded) models of the sort described previously for the colonization of the New World and the Pacific should not be viewed as the last word on this subject—instead, they are starting points for additional investigations to unravel the finer details of what really happened in the past.

CHAPTER 17

SPECIES-WIDE SELECTION

In this and the following chapter, we will consider the impact of what most people would consider to be the most significant of the various forces that have influenced human evolution, namely selection. After all, selection lies at the core of Darwinian evolution, involving increases in reproductive fitness over time via new genetic adaptations, often in response to some change in circumstances—climate, parasites, disease, new sources of nutrition, and so forth. So, how do we go about detecting genes that have been influenced by selection, and how do we figure out the evolutionary reason for a particular genetic adaptation? These are the questions we will address in this and the subsequent chapter.

We will start by making a distinction between selection that happened prior to the origin of modern humans and selection that occurred after modern humans began spreading around the globe. The first type of selection thus resulted in genetic changes shared by all modern humans (and hence can be called **species-wide selection**) and is the subject of this chapter. In the next chapter, we will consider selection that has resulted in genetic changes shared by only a subset of modern humans (and hence is often referred to as **local selection**). The reason why we consider these separately is because (mostly) different methods are used to detect these different types of selection; moreover, the resulting implications concerning human evolution also differ for these two types of selection. The genetic changes that were selected for after our lineage diverged from that of chimpanzees can be thought of as those changes that made us human, while the genetic changes that were selected for as modern humans spread across and out of Africa can be thought of as those changes that enabled us to colonize more of the globe, and a wider variety of environments, than any other species.

In general, there are two types of approaches that can be used to try to identify genes that have been subject to selection. The first is referred to as the **candidate gene** approach and involves starting with a phenotype of interest, coming up with a list of genes that might influence that phenotype (these are the candidate genes), and then analyzing the variation in these genes to see whether any of them show a signal of selection (more on this later, as the signal one looks for depends on whether one is looking for species-wide or local selection). This approach has a long history in genetics, was originally developed for identifying genes involved in diseases, and has produced a number of success stories. However, it does suffer from the following drawback: while it is true that if you've been very clever (or very lucky) in your choice of candidate genes you can end up with a nice story, it is also the case that if you haven't been so clever or lucky in your choice of candidate genes and none of them shows any evidence of selection, then you end up doing a lot of work with nothing to show for it. Therefore, most studies of selection nowadays use the second approach, which involves scanning genome-wide data for unusual features that might indicate the impact of selection, analogous to the use of genome-wide association studies (GWAS, mentioned previously in Chapter 7) as a tool for identifying candidate genes for complex diseases. In a typical GWAS, you genotype a few hundred thousand or million SNPs (using one of the commercially available SNP chips) in people with and without the disease (cases and controls) and look for SNPs that are associated with the disease. With this approach, you find any gene that might be associated with the disease, not just genes you already suspect. Similarly, in a genome scan for selection, you screen variation across the genome in your set of samples and look for unusual patterns of variation that suggest that selection has influenced that particular region of the genome. This approach thus, in principle, finds any signature of selection in the genome, not just those that you have some reason for suspecting. With this

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

in mind, we will consider the signals of species-wide selection in genomic data first, followed by signals of local selection in the next chapter.

SPECIES-WIDE SELECTION

The overall goal of the institute I work at (the Max Planck Institute for Evolutionary Anthropology) is to investigate what makes humans human, from the standpoint of genetics, paleontology, archaeology, primatology, linguistics, and psychology. You might think that with the determination of the human and chimpanzee genome sequences (in 2001 and 2005, respectively), for those of us in genetics, our work was over. After all, the genome sequences provide a complete catalog of all of the genetic differences between humans and our nearest living relatives, chimpanzees, so the genetic changes that made us what we are must be somewhere among those genetic differences. Unfortunately, it's not that simple (or rather fortunately, for those of us who want to keep our jobs!), as there are some 20 million or so genetic differences between humans and chimpanzees, and (so far) we have not been able to sort out those that were important in the evolution of our lineage from those that reflect random genetic changes without any selective impact (not to mention those that were important in the evolution of the chimpanzee lineage). Predicting the functional impact of a mutation remains one of the biggest challenges and not just in evolutionary genetics—nowadays, it is becoming increasingly common to sequence the exome (all of the protein-coding genes) or even the entire genome from individuals with a disease that is suspected to have a genetic basis, in order to identify the disease-causing mutation(s). The problem here is similar—how do you determine which of the many mutations revealed by exome/genome sequencing are truly associated with the disease?

With disease mutations, there are various tricks one can use to help narrow the search, such as comparing family members with and without the disease, or comparing several unrelated individuals with the disease for shared mutations (or for different mutations in the same gene, which would then implicate that gene in the disease). And it is not like we have absolutely no clue about the potential functional impact of a mutation—recall from Chapter 2 that a nonsense or frameshift mutation is likely to profoundly disrupt the function of a gene, and nonsynonymous mutations are also potential candidates for having big effects on gene function.

So in an evolutionary context, we can similarly look for those types of mutations that are more likely to influence gene function. But we can also do more. By definition, those genetic changes that were important in making humans human were selected for during our evolution and hence should show the signature of Darwinian natural selection. So if we know what sort of footprint natural selection leaves behind in the genome, we can search genomic data for such footprints and use this information to help identify the truly important genetic changes. It turns out that there are many potential signatures of selection in the genome and more are being identified all the time—too many to cover exhaustively. So, what we will instead do in the remainder of this chapter is discuss a few of the main signatures of selection in the genome, how to detect them, and then go through some examples of what we have learned.

NONSYNONYMOUS MUTATIONS AND THE dN/dS RATIO

Nonsynonymous substitutions (those that alter the amino acid sequence of the encoded protein) are often a focus of selection studies. This is not so much because they are necessarily the most important kind of mutation when considering adaptive evolution but rather because it is easy to pick them out—recall from Chapter 6 that there is good reason to think that mutations that influence gene regulation are just as important evolutionarily, if not more so, but they are a lot harder to identify. Also, at least some nonsynonymous substitutions are likely to have functional effects. So, one footprint of positive selection is a gene with an excess number of nonsynonymous substitutions. However, genes with higher than average mutation rates will also have more nonsynonymous substitutions than the average gene. So how can we tell whether an excess number of nonsynonymous mutations is due to selection or to a high mutation rate? Simple: selection influences only nonsynonymous substitutions, while the mutation rate influences all positions in the gene, so to correct for mutation rate differences, what you do is divide the number of nonsynonymous substitutions by the number of synonymous substitutions (i.e., those mutations in a protein-coding gene that do not alter the amino acid sequence of the protein). The idea is that if a gene simply has a higher than average mutation rate, then both nonsynonymous and synonymous substitutions will be elevated, and by taking the ratio you cancel out the effect of a higher mutation rate. But if only the nonsynonymous substitutions are elevated for a particular gene, because of selection, then the ratio of nonsynonymous to synonymous substitutions will be bigger for such genes than for genes that have not been subject to such selection. You also have to correct for the number of potential mutations in a gene that will result in a nonsynonymous substitution versus those that will result in a synonymous

substitution; the overall result is called the **dN/dS** ratio.

The bigger the dN/dS ratio for a gene, the more nonsynonymous substitutions (relative to synonymous substitutions) that gene has experienced. In general, genes that are evolving completely under neutrality, with no constraints whatsoever on the kinds of mutations that can occur, are expected to have a dN/dS ratio of 1. In practice, such dN/dS ratios are, with very few exceptions, observed only for pseudogenes (gene copies that no longer produce a protein), as practically every functioning gene evolves under some sort of constraint. Hopefully, it makes sense that some mutations simply cannot be tolerated as they disrupt some vital function of the gene. So, the vast majority of genes have dN/dS ratios less than 1, and the smaller the dN/dS ratio, the greater the functional constraint on the gene (i.e., the fewer the amino acid substitutions that are compatible with gene function). However, genes that have experienced positive selection for repeated amino acid substitutions will have dN/dS ratios bigger than 1. Note that just one nonsynonymous substitution probably won't result in an elevated dN/dS ratio; it takes either several amino acid substitutions that were selected for at the same time or selection for repeated amino acid substitutions over a period of time to produce a dN/dS ratio bigger than 1. It should therefore come as no surprise that genes that have dN/dS ratios bigger than 1 are usually involved in the immune response to infectious disease. Pathogens evolve much more rapidly than we do because of their much shorter generation time (often measured in hours, not years) and are hence constantly changing their surface characteristics (i.e., antigens), so there is strong selection for those components of our immune system that recognize and bind to these pathogen surface characteristics to change accordingly. And once we have evolved resistance to one form of a pathogen, there is then selection for the pathogen to change to a new form that evades our immune response—it is an evolutionary "arms race" where we are inevitably one step behind.

It might seem that there is only limited information to be gained from analyzing dN/dS ratios, as multiple amino acid substitutions are required to produce a dN/dS ratio bigger than 1, so any selection involving just one or two amino acid substitutions won't be detected. But we can do better than this, because we can also analyze dN/dS ratios in the context of a phylogenetic tree and look for branches in the tree that have experienced even a modest increase in the dN/dS ratio for a gene (even if the overall dN/dS ratio does not exceed 1). For example, we can examine dN/dS ratios in a phylogenetic comparison of humans, chimpanzees, and an outgroup (e.g., mice) and look for genes that show a significant increase in the dN/dS ratio on the human lineage. Let's go through an example of a gene for which a signature of selection was detected by this approach and then see what we have learned subsequently about this gene.

In 2001, it was reported that several members of a family with verbal dyspraxia (deficits in the orofacial muscular movements necessary for the production of speech) and with problems with some aspects of acquiring and using language all shared a novel nonsynonymous mutation in a transcription factor gene called FOXP2 (Lai et al. 2001). An unrelated individual with a similar phenotype had a translocation (chromosome rearrangement) that disrupted the FOXP2 gene; taken together, these two observations strongly implicate FOXP2 as a gene important in the acquisition and production of language. Given the undisputed importance of language to humans, understanding how language evolved is obviously a central question in human evolution, and it appeared that studying the evolution of FOXP2 might provide a way to approach this question. Therefore, Wolfgang (Wolfi) Enard, then a graduate student with paleogeneticist Svante Pääbo, began investigating FOXP2. He started by comparing the human FOXP2 sequence to the mouse FOXP2 sequence, thinking that if FOXP2 had indeed been selected for during human evolution, then human FOXP2 might show lots of nonsynonymous differences when compared with mouse FOXP2. Disappointingly, there were only three nonsynonymous differences between human and mouse FOXP2, which is actually less than the average difference between mouse and human proteins. In fact, FOXP2 is in the top 5% of genes in terms of conservation (i.e., similarity) of amino acid sequences between humans and mice. It therefore seemed that there was strong selection to conserve FOXP2 function during mammalian evolution (at least, that part of mammalian evolution covered by humans and mice) and little indication of any accelerated change in human FOXP2. Fortunately, Wolfi persevered by determining the chimpanzee FOXP2 sequence (as well as the gorilla, orangutan, and rhesus monkey FOXP2 sequences), and now things started to get interesting (Enard et al. 2002): chimpanzee FOXP2 differs from mouse FOXP2 by just one amino acid and from human FOXP2 by two amino acids (Figure 17.1). In other words, during the 130 or so million years of evolution that separate mice from the common ancestor of humans and chimpanzees, there was just one amino acid substitution in FOXP2. Then, during the 6 million years or so of evolution after the divergence of the human and chimpanzee lineages, there were two amino acid substitutions in human FOXP2. Two amino acid substitutions may not sound like much, but in the overall phylogenetic context of FOXP2, it does represent a significant acceleration in the rate of FOXP2 evolution

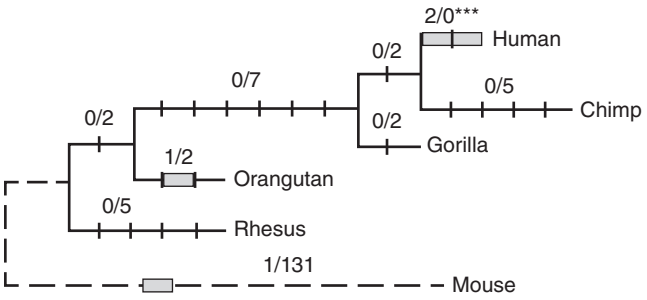


FIGURE 17.1

Phylogenetic tree relating FOXP2 sequences from different primates, rooted with the mouse FOXP2 sequence. Bars indicated nucleotide changes, with gray bars indicating amino acid changes. The first number on each branch indicates amino acid substitutions while the second number indicates silent substitutions; the asterisks indicate a significant difference between humans and chimpanzees. Note that the tree is based on a single sequence from each species; additional sequences from humans do reveal additional variation, albeit all humans share the same two amino acid substitutions. Reprinted with permission from Enard, W., et al., "Molecular evolution of FOXP2, a gene involved in speech and language," *Nature* 418:869, 2002.

(Figure 17.1). Moreover, Wolfi sequenced the *FOXP2* gene from a worldwide sample of humans and found that all humans share the two nonsynonymous differences with respect to chimpanzee FOXP2, suggesting that these are indeed fixed differences between humans and chimpanzees.

It therefore becomes of interest to know when these two amino acid substitutions became fixed in the human population, as this would give some indication as to whether language was a relatively old or a relatively recent development in human evolution (assuming, of course, that these amino acid substitutions in FOXP2 do influence human language skills, which was still a big if). Population geneticist Molly Przeworski (then a postdoctoral fellow with Svante Pääbo) used a coalescent simulation approach and the sequence data from the worldwide sample of humans to estimate the time since fixation for these two mutations, based on the amount of variation in the region of FOXP2 surrounding them. Assuming that fixation happened relatively quickly, which is a reasonable assumption for strong selection, then all of the variation observed today accumulated after fixation and hence can be used to estimate how much time has elapsed since fixation. The resulting estimate was 0–200,000 years ago (Enard et al. 2002), suggesting that the fixation of these two amino acids (and by implication, any associated effect on language skills) was a relatively recent development in human evolution.

Unfortunately, it is easy to forget that such estimates come with lots of assumptions that can have a big impact on the results. In this case, if the upper limit of 200,000 years ago for the fixation of the human FOXP2 alleles at these two amino acid positions is indeed correct, then we would predict that Neandertals and Denisovans should have the chimpanzee (ancestral) alleles at these two positions—which would also suggest that Neandertals and Denisovans lacked whatever changes in language abilities were associated with the human form of FOXP2. It therefore was a source of some consternation when a targeted study of FOXP2 in Neandertals showed that the Neandertal FOXP2 sequence carried the derived human alleles at these two positions (Krause et al. 2007). That this result was authentic and did not reflect human contamination was confirmed by the subsequent Neandertal and Denisovan genome sequences: it is clear that both Neandertals and Denisovans have FOXP2 amino acid sequences that are identical to human FOXP2 (Green et al. 2010; Reich et al. 2010).

So what went wrong with the estimate for the age of the fixation of the two amino acid substitutions? The problem seems to be not with the method itself but rather with the assumption that the signal of selection on FOXP2 that is being dated is actually associated with the fixation of these two amino acids. It now appears that the selection event that is being dated is related to some other change in FOXP2, not to the two amino acid substitutions. The reason for thinking this is too complicated to go into here but has to do with the patterns of linkage disequilibrium (associations among the alleles at linked polymorphic sites) in and around these two amino acid substitutions; the results indicate that a different region of FOXP2 was the likely target of selection (Ptak et al. 2009). Indeed, in recent work, Tomislav Maricic from Svante Pääbo's group has found a mutation in an intron of FOXP2 that is in the putative selection region and that influences the expression of FOXP2 (Maricic et al. 2013). This mutation is present at ~98% frequency in a worldwide sample of modern humans, but Neandertals and Denisovans have the ancestral allele. So, this is a promising candidate for the signal of recent selection on FOXP2 in modern humans, but more work is needed to verify whether this is really the case—for example, it would be interesting to know whether the ~2% of modern humans with the ancestral allele have any associated phenotype related to language acquisition/production.

Meanwhile, what about the two amino acid substitutions? Even though they may not be the target of the signal of recent selection on FOXP2 in humans, still, the occurrence of two amino acid substitutions on the human lineage does represent a significant acceleration in the rate of FOXP2 evolution, so it would be interesting to know what effect—if any—they might

have on the function of FOXP2. So, how might we go about investigating the functional effects of these mutations? If we were talking about bacteria or fruit flies, then we could engineer strains of bacteria or fruit flies that differed by these two amino acid substitutions at FOXP2 but were otherwise genetically identical, and thereby see what effect the amino acid substitutions had. But with humans, our options are rather limited. It is neither ethical nor practical to create humans who lack these two FOXP2 amino acid substitutions, and the same holds for introducing these two FOXP2 amino acid substitutions into chimpanzees; such genetic engineering of humans or chimpanzees is out of the question. We could see what happens if we manipulate the *FOXP2* gene in human or chimpanzee cells that are growing in the laboratory—and we will see an example in the next chapter in which manipulating and studying cells in the laboratory was informative—but it is not clear how we could learn anything about complex behaviors (like language acquisition) from studying cells in the laboratory.

There is one other approach to consider and that is to make a humanized mouse model—that is, introduce these two amino acid substitutions into an inbred line of mice and see what happens (in the best case scenario, the mice start talking!). This is an audacious experiment to carry out, because it is expensive and takes a long time (a good year or so) to make a humanized mouse, and there is no guarantee that you'll learn anything useful—mice and humans do share a lot of biology in common since they are both mammals, but there are some rather obvious and significant differences as well. In the worst-case scenario, you go to a lot of time and expense and don't learn anything. Still, FOXP2 was a good candidate for making a humanized mouse because both of the amino acid substitutions of interest are in the same exon (which technically makes it easier to introduce them into mice), and the FOXP2 protein is otherwise essentially identical between humans and mice (with just one additional amino acid difference), so the FOXP2 of nonhumanized mice is basically like chimpanzee FOXP2—no need to make a "chimpanized" mouse as well to compare the effects of human versus chimpanzee FOXP2. So, Wolfi Enard decided to take the plunge and had a humanized FOXP2 mouse created, and fortunately the gamble paid off (Enard et al. 2009). The humanized mice were subjected to a large battery of tests that covered every aspect of mouse morphology, physiology, anatomy, and behavior that you could think of (plus lots of things you would never think of). And the result? Not talking mice, alas—but when mice pups are removed from the nest, they emit ultrasonic vocalizations, which presumably help the mother find them and return them to the nest, and interestingly, the ultrasonic vocalizations of humanized FOXP2 mice differ from those of normal mice. Moreover, humanized FOXP2 mice show alterations in exploratory behavior and in the chemistry and circuitry of the basal ganglia. The basal ganglia is a part of the brain that is implicated in language acquisition in humans as well as in the acquisition of songs by birds, so to see effects on this part of the brain in the humanized FOXP2 mice was quite a promising result.

However, one issue that arises in general with humanized mice is that the effects one sees may simply be due to incompatibilities between the humanized version of the protein and the mouse proteins that it interacts with. In other words, the phenotypic differences associated with the humanized FOXP2 mice may reflect the fact that the human version of FOXP2 just doesn't work as well in a mouse as mouse FOXP2 does. To control for this possibility, Wolfi also investigated a mouse model that was heterozygous for a mutation that had been similarly engineered but that knocks out FOXP2 activity (homozygotes for this mutation die before birth, as FOXP2 function is essential for life). The idea is that if the phenotypic effects associated with the humanized FOXP2 mouse are actually due to reduced activity of FOXP2 because of incompatibilities associated with human FOXP2 in mice, then mice with a mutation that reduces endogenous mouse FOXP2 activity should show similar phenotypic effects. In fact, mice with reduced FOXP2 activity showed phenotypic effects that were in the opposite direction of those exhibited by humanized FOXP2 mice.

So, it is likely that the phenotypic effects seen in the humanized FOXP2 mice are indeed due to the two amino acid substitutions. This, in turn, suggests that these amino acid substitutions may have played a role in the evolution of human language by altering the chemistry and circuitry of the basal ganglia, perhaps fine-tuning this part of the brain to be more receptive to language acquisition. After fixation of these amino acid substitutions during human evolution, there may have been further selection on the expression of FOXP2, perhaps in accord with its new role in language and the increasing importance of language to humans, to account for the signature of a recent selective sweep. At any rate, this makes a nice (if speculative) story, and both the humanized FOXP2 mouse model and some of the specific questions raised by this model are currently being actively investigated—for example, since FOXP2 is a transcription factor involved in activating the expression of other genes, it is of interest to know what the other genes are that are activated, and how gene expression differs between the human and chimpanzee forms of FOXP2. And keep in mind, this entire line of research started with the simple observation of an accelerated rate of nonsynonymous substitutions for FOXP2 on the human lineage.

TESTS BASED ON THE ALLELE FREQUENCY DISTRIBUTION

An acceleration in amino acid substitutions is one signature of natural selection. Another signature is based on expectations concerning the allele frequency distribution expected under neutrality versus that expected under selection, and several such tests for neutrality have been developed. One of the first, and one that is still widely used, was developed by the population geneticist Fumio Tajima (Tajima 1989) and is known as Tajima's D test. The basis of the test is quite simple. Suppose we have a sample of DNA sequences from a population. Assume that every mutation that has occurred in our sample happened at a different position in the sequence (this is the so-called "infinite sites" model); this is an important assumption because it means that every mutation that has occurred can be detected. Then, the observed average heterozygosity per nucleotide position (called π) provides an estimate of the parameter $\Theta = 4N_e\mu$ (where N_e is the effective population size and μ is the neutral mutation rate—see Chapter 5 for a refresher on why Θ is of interest). Let's call this $E(\pi)$, for the expected value of Θ based on π. Similarly, the number of polymorphic sites (called k) in our sample of sequences can also be used to estimate Θ, so let's call this $E(k)$. Clearly, under neutrality, the difference between the two estimates, $D = E(\pi) - E(k)$, should be zero. But if selection has influenced the variation in our sample of sequences, then it turns out that D will be different from zero, because different types of selection will affect π and k differently. As depicted in Figure 17.2, positive selection will reduce variation around a selected variant, because the selected variant increases in frequency rapidly until it reaches fixation—this is also known as a **selective sweep**. Hence, most of the variation in a selective sweep region

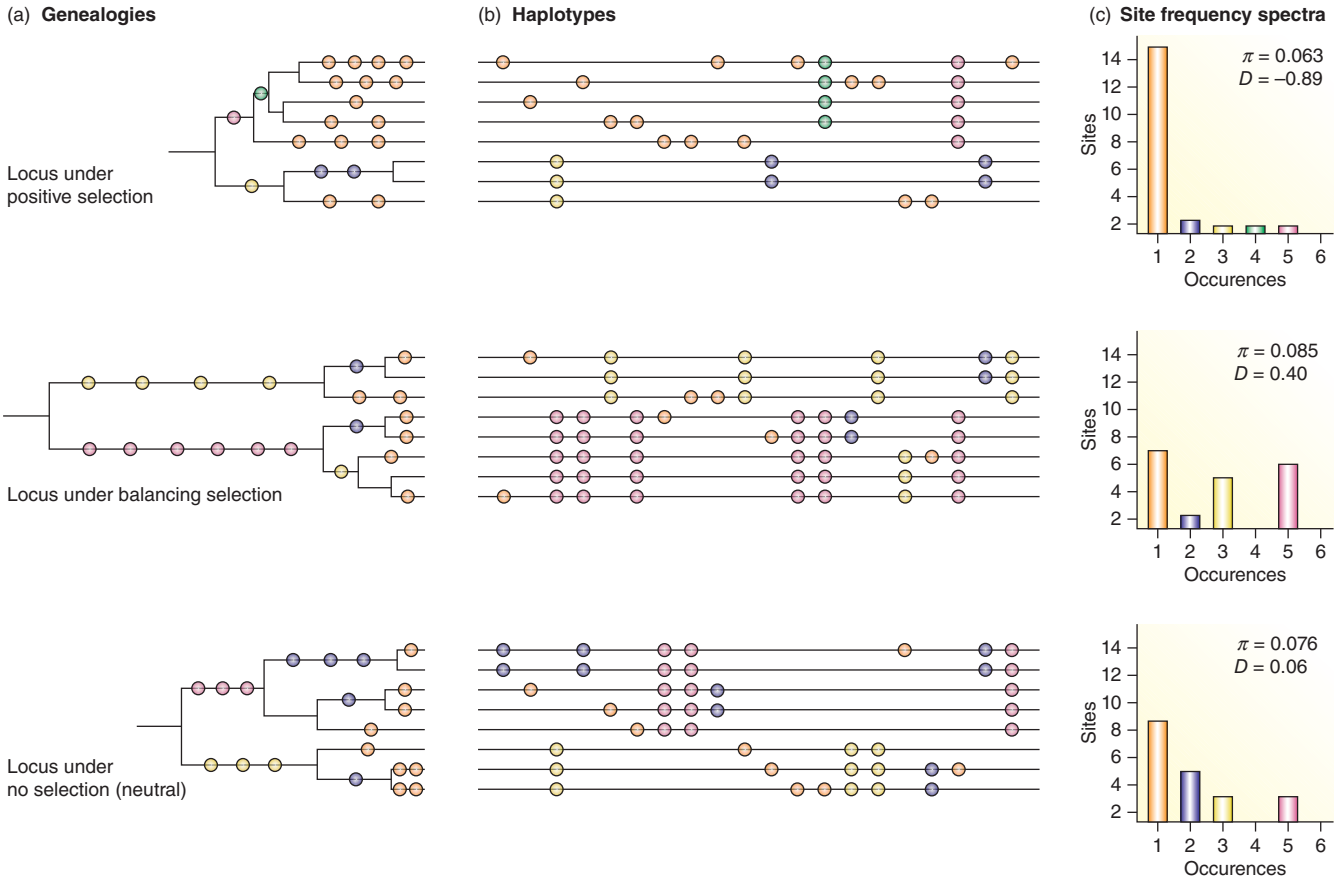


FIGURE 17.2

The expected effects of positive selection, balancing selection, and neutrality on the genealogy (phylogeny) of a locus, the distribution of haplotypes, the site frequency spectrum, and estimates of π (heterozygosity) and Tajima's D. Note that the genealogy for a locus under positive selection also looks like what we would expect for a population expansion, and the genealogy for a locus under balancing selection looks like what we would expect for a bottleneck—this is important later. Reprinted with permission from Bamshad, M., and Wooding, S.P., "Signatures of natural selection in the human genome," *Nature Reviews Genetics* 4:99, 2003.

will consist of low-frequency variants that were either introduced by the occasional recombination event during the selective sweep or reflect new mutations that occurred during or after the selective sweep. Such low-frequency variants contribute more to k (the number of polymorphic sites) than to π (average heterozygosity); a variant at a frequency of 1% counts just as much toward k as a variant at a frequency of 50%, but the former has a heterozygosity of just 2% while the latter has a heterozygosity of 50%. The overall result is that $E(k)$ is bigger than $E(\pi)$, and so D is less than zero.

Suppose that instead of positive selection, our DNA sequences come from a region of the genome that has been influenced by balancing selection (i.e., selection for heterozygosity). Then, as also depicted in Figure 17.2, there will be an excess of heterozygosity (π) relative to the number of polymorphic sites (k), and so $E(\pi)$ will be bigger than $E(k)$, resulting in a value of D that is bigger than zero.

So, to summarize, negative values of D that are significantly different from zero indicate positive selection, while significant positive values of D are indicative of balancing selection; values of D that do not differ significantly from zero are consistent with no selection (neutrality). And what do we get when we calculate Tajima's D from human DNA sequence data? Figure 17.3 shows the distribution of D values for 313 genes, sequenced in a diverse sample of 82 humans. The majority of the D values are significantly less than zero, and the average D value is −0.97. So what is going on here—does this mean that the majority of human genes have experienced positive selection? This would certainly reinforce the view that humans are special creatures after all, as it apparently took lots of positive selection on lots of genes for us to evolve! Alas, for those who want to see evidence of our specialness in our evolution, there is a much more prosaic explanation for all of these negative D values. It turns out that, as with many such tests, Tajima's D has a number of underlying assumptions, and significant results may reflect a violation of one of these assumptions, rather than a failure of the condition that one thinks is being tested (e.g., neutrality vs. selection). For Tajima's D, one of the underlying assumptions is that the population has been constant in size over time. But, as we saw back in Chapter 12, changes in population size will influence the allele frequency distribution (indeed, this is the reason why we can use current patterns of genetic variation to investigate the past demographic history of a population). In particular, a population expansion will result in more rare alleles than expected in a population that has been constant over time—exactly the same result expected with positive selection. So, a population expansion will result in a negative value for Tajima's D, and if we fit a model of population expansion to the genetic data used to produce the distribution of Tajima D values shown in Figure 17.3, then the expected value for Tajima's D is indeed negative, and the vast majority of the observed D values do not differ significantly from that expected with a population expansion.

In general, how can we tell whether a negative Tajima's D value (or other significant test statistic) reflects selection or population expansion? Simple: population expansions (or any other demographic process) are expected to have the same average

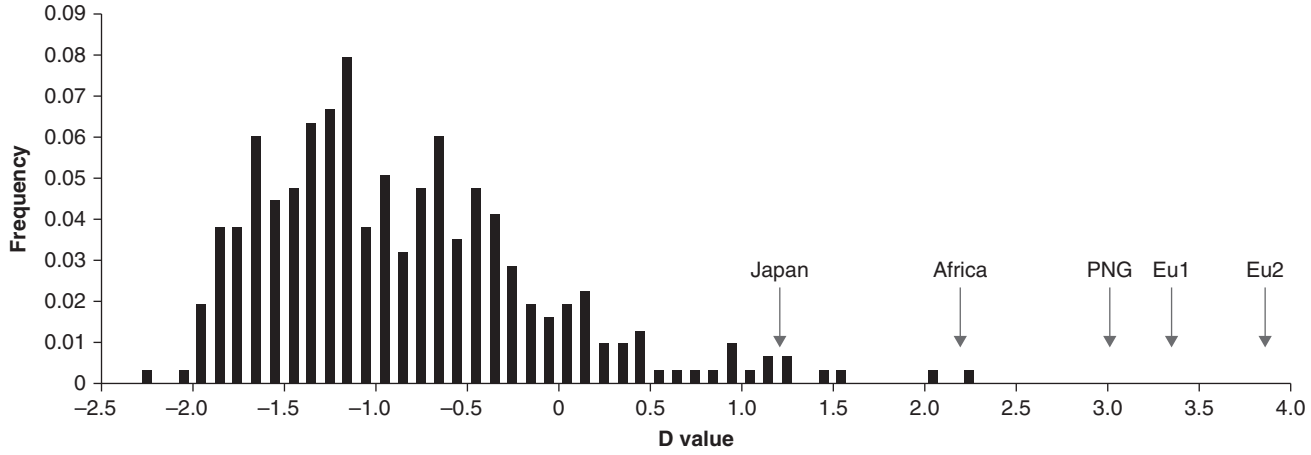


FIGURE 17.3

Distribution of Tajima's D values for 313 genes that were sequenced in 82 individuals. The average D value is −0.97, which is significantly different from zero (the value expected under neutrality). Also indicated are the Tajima D values observed for the prion protein gene in several human populations. Reprinted with permission from Stoneking, M., "Widespread prehistoric human cannibalism: Easier to swallow?," *Trends in Ecology & Evolution* 18:489, 2003.

effect across the entire genome, whereas selection is expected to influence at most a few specific genes. So, if we sample a lot of genes (or genomic regions) randomly and find that the average D value is negative, then we can be reasonably certain that demography is responsible. Genes that are nonetheless outliers in the overall distribution of D values are then candidates for selection, as we shall see in a moment. The take-home message: a significant result in any statistical test reflects failure for any of the assumptions of the test to hold, so you need to understand all of the specific assumptions of the test in order to properly interpret the results. Otherwise, you may not be testing the aspect of the data that you think you are testing. In the case of Tajima's D, two of the crucial assumptions are that the gene is neutral and that the population size has been constant, so a significant result could reflect failure of either of these assumptions (or both). And in case you were wondering, these are not the only crucial assumptions of Tajima's D; we'll now turn to a specific example involving Tajima's D that nicely illustrates this point.

In 2003, a provocative study appeared of variation in one particular gene that suggested that our ancestors regularly ate one another—that is, that cannibalism was routinely and widely practiced by our ancestors (Mead et al. 2003). How does one arrive at a conclusion of widespread prehistoric cannibalism from such a study of genetic variation? The gene in question encodes a protein called the prion protein, a protein found on cell membranes whose function is still unclear. However, the prion protein can sometimes assume an altered conformational state that leads it to aggregate and form plaques in the brain and other tissues, which can then lead to neurodegenerative, "prion" diseases such as Creutzfeld–Jakob disease in humans, or scrapie in sheep. How and why this change in conformational state comes about is still not clear; it appears to involve some sort of spontaneous misfolding of the protein. Interestingly, misfolded prion proteins can cause "normal" prion proteins to also misfold (and how this happens is also not known), and so prion diseases can be transmitted from one individual to another by consuming infected tissues. The idea that the misfolded prion protein alone is sufficient to transmit the prion disease from one individual to another was quite controversial when it was first proposed by the neurologist Stanley Prusiner in 1982 (Prusiner 1982), in which he coined the term "prion" for a proteinaceous infectious particle—that is, an infectious particle that consisted entirely of proteins. Standard dogma held that only nucleic acids (DNA or RNA) could be infectious (because only nucleic acids can replicate—proteins cannot), so Prusiner was ridiculed for daring to propose that a protein particle, without any accompanying nucleic acid, could be infectious. The obvious answer—according to his critics—was that there was a small amount of nucleic acid present in the infectious particles that his assays were not able to detect. Prusiner refused to be swayed, despite criticism that went beyond the science involved and attacked him personally, and despite difficulties in obtaining funding for his work. Gradually, most of the skeptics were converted when increasingly sophisticated assays still failed to turn up any hint of nucleic acid. While not everyone agrees that prions consist only of proteins, this is now the mainstream view, and Prusiner was rewarded for his perseverance with a Nobel Prize in 1997.

The importance of prion diseases was amply demonstrated in the late 2000s with the realization that bovine spongiform encephalopathy (BSE), a prion disease of cattle with a lengthy incubation time of up to 8 years, could be transmitted to humans via the consumption of infected tissues (mostly by eating beef contaminated with brain or spinal cord tissue, although BSE prions can be found in just about any tissue in infected individuals). Given the very lengthy period of time between the exposure and the onset of disease, it was initially feared that when the first cases in humans turned up, this could be just the beginning of a potentially catastrophic epidemic of neurodegenerative disease in humans. However, the number of cases of BSE in humans has leveled off at "only" about 200 deaths, and with massive culling of infected animals and strict controls on feeding offal (entrails and other yummy bits) to cattle (which is the primary means by which BSE spreads in cattle), it seems that the potential for disaster has been avoided. Still, something to think about the next time you have an urge for a Big Mac!

Another example of a prion-associated neurodegenerative disease is kuru, which is almost entirely restricted to the Fore people of highland New Guinea. Because kuru tends to run in families, it was thought to be a genetic disease, but the virologist Carleton Gajdusek won a Nobel Prize in 1976 for showing that it was actually an infectious disease. The Fore used to practice ritual cannibalism (before Europeans halted the practice), in which relatives and friends of a dead individual would consume the remains, as a sign of respect. Kuru was widespread among the Fore when Gajdusek started working in the area in 1957. Gajdusek thought that a slow virus (one with a long incubation time of several years) was responsible, and that the virus was transmitted by eating the brains of victims (which were reserved for the nearest relatives of the deceased, accounting for the familial transmission of the disease). He showed that the disease could be transmitted to chimpanzees and had a long incubation time, which was the first demonstration that a neurodegenerative disease in humans could

be infectious—even though he was wrong about the nature of the infectious agent, as it turns out that prions are responsible, not a virus. Incidentally, Gajdusek used his Nobel Prize money to set up a foundation to bring boys from New Guinea to the United States for schooling, an apparently laudable use of the money, but was later convicted of sexually molesting some of the boys—yet another example of how not all Nobel Prize winners are of sterling character.

Anyway, previous work had shown that a nonsynonymous polymorphism at codon 129 of the prion protein gene was associated with susceptibility to sporadic neurodegenerative diseases and to BSE (Palmer et al. 1991). Homozygotes for either valine or methionine at this codon are more susceptible than heterozygotes; why this is the case is not known, but apparently heterozygosity at codon 129 (which we will call 129het) inhibits the formation of the altered prion protein conformation that leads to disease. Simon Mead, who carries out research on various neurodegenerative diseases, wondered whether 129het had any effect on kuru, so he led a study in which 30 Fore were genotyped for this polymorphism (Mead et al. 2003). All of them were older than 50 years and had participated in numerous mortuary feasts, so all had ample exposure to kuru but so far had not developed any signs of the disease. Astonishingly, 23 of them (77%) were 129hets, which is significantly different both from Hardy–Weinberg expectations (i.e., the expected genotype frequencies if there was no selection or other evolutionary force acting on this polymorphism—see Chapter 4 for a refresher on Hardy–Weinberg) and from a sample of Fore that had never participated in mortuary feasts. Apparently, homozygotes for position 129 were susceptible to kuru and hence had contracted kuru and died after participating in mortuary feasts, resulting in the excess heterozygosity at position 129 among those who still survived despite exposure to kuru. Thus, balancing selection has been acting to favor 129het over homozygosity for this amino acid position, presumably because 129hets are protected from kuru.

If balancing selection is operating on the prion protein gene in the Fore, then Tajima's D should be significantly positive (as discussed previously), and indeed it is. But in addition to the Fore, Mead and colleagues analyzed the prion protein gene in four additional populations (one from Africa, one from Japan, and two from Europe), and here is where it gets interesting: all four populations showed significantly positive values for D, suggesting that all human populations have experienced balancing selection on the prion protein gene. A rough estimate for the age of the polymorphism (based on the amount of flanking variation) is about 500,000 years, suggesting that the signal of balancing selection could also be that old. Mead and colleagues considered several possible explanations for the signal of balancing selection, including resistance to some (unknown) infectious disease, or protection against prion diseases carried by animals that our ancestors regularly consumed. However, the explanation they favored was that "… repeated episodes of endocannibalism-related prion disease epidemics in ancient human populations made coding heterozygosity at PRNP (the prion protein gene) a significant selective advantage leading to the signature of balancing selection observed today." In other words, our ancestors liked to eat one another—and I have to confess that I rather liked this study and wrote about it in a commentary for the journal *Trends in Ecology and Evolution* (Stoneking 2003).

Alas, as T.H. Huxley said in a presidential address to the British Association for the Advancement of Science in 1870, the great tragedy of science is the slaying of a beautiful hypothesis by an ugly fact; and soon after the study was published, others pointed out a problem with the data used to calculate the values for Tajima's D. In order for the neutrality test based on Tajima's D to be valid, you need accurate, unbiased estimates of the number of polymorphic sites and of the heterozygosity at each site for your gene of interest. Ordinarily, this is easily accomplished with DNA sequence data; however, rather than sequencing the prion protein gene, Mead and colleagues decided to save time and money (this was back when DNA sequencing was more time-consuming and expensive) and simply genotype informative SNPs in the prion protein gene. And how do you know which are the informative SNPs? You select SNPs which, from previous studies, are known to be polymorphic in humans. And there is the rub—as discussed back in Chapter 7, this sort of procedure introduces an ascertainment bias. The SNPs that you end up analyzing are more heterozygous than the SNPs that you don't analyze (because, by definition, you don't analyze any rare SNPs that weren't previously detected), and so you end up overestimating the average heterozygosity for your gene. This, in turn, leads to values for Tajima's D that are skewed in the direction of being too positive. Indeed, a later study of the prion protein gene that carried out DNA sequencing (Soldevila et al. 2006) provided convincing evidence that the ascertainment bias was indeed responsible for the significantly positive Tajima's D values (Figure 17.4); the sequence-based values for Tajima's D are actually negative and fall within the distribution observed for other genes (Figure 17.3). So, while the evidence for balancing selection on codon 129 specifically in the Fore remains compelling, there is no evidence for balancing selection on the prion protein gene in any other human population.

This is not to say that there is no evidence for cannibalism by our ancestors—cannibalism is one of those

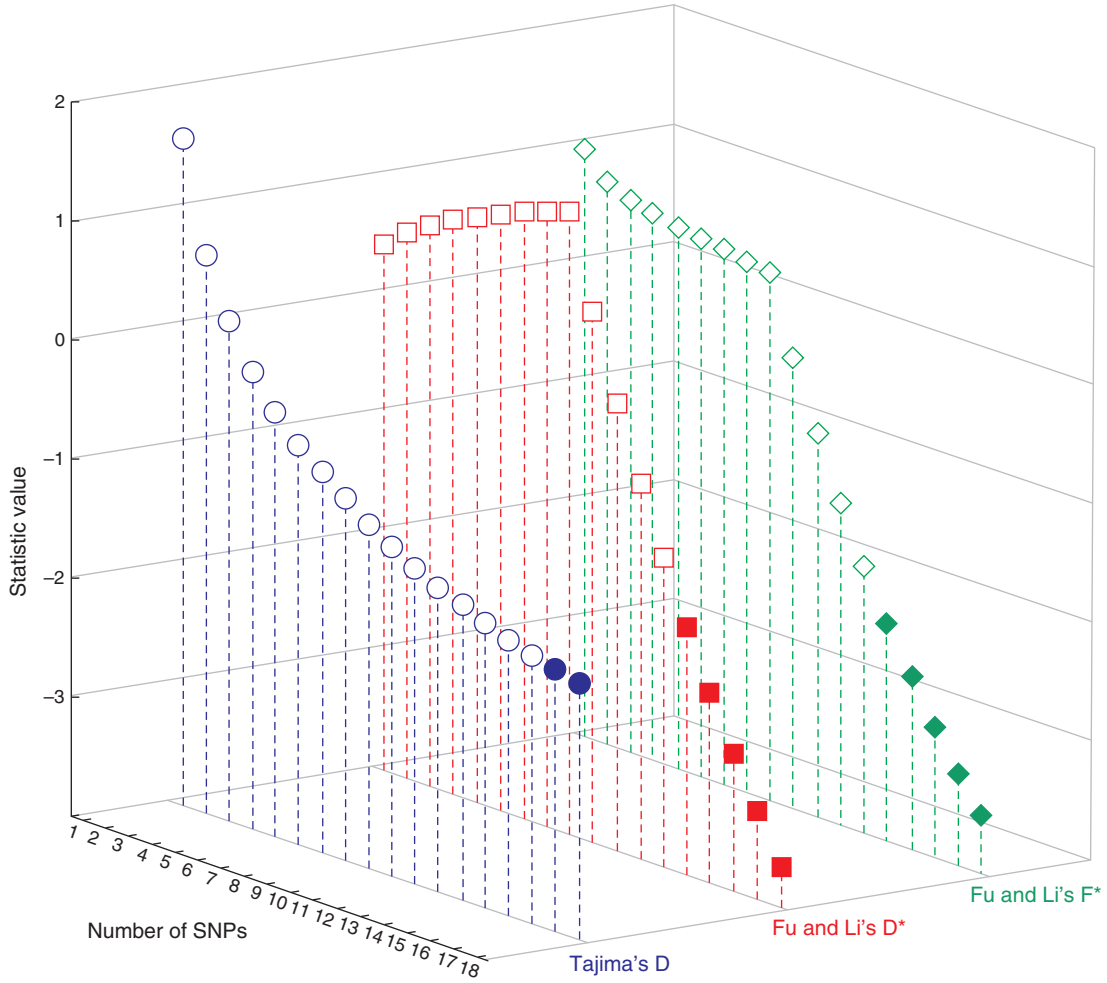


FIGURE 17.4

Changes in the value of Tajima's D (and other neutrality parameters) as more SNPs from the prion protein gene are added. A full sequence analysis of the prion protein gene found 18 polymorphic sites; the values of Tajima's D are shown as sites are added sequentially in decreasing order of heterozygosity. As SNPs with lower heterozygosity are added, Tajima's D goes down, until with all of the SNPs included a significant negative value (indicated by solid circles) for Tajima's D is obtained. Genotyping only the most heterozygous SNPs, as Mead and coworkers did, results in an inflated positive value for Tajima's D because of the ascertainment bias; full sequence analysis results in a negative value for Tajima's D that does not differ from the genomewide average shown in Figure 17.3. Reprinted with permission from Soldevila, M., et al., "The prion protein gene in humans revisited: lessons from a worldwide resequencing study," *Genome Research* 16:231, 2006.

nasty human behaviors that many people would like to deny on the grounds that it isn't very nice to think that humans would ever do that, and therefore the bar for demonstrating cannibalism has been set very high. For example, some anthropologists discount any anecdotal observations of cannibalism in other societies—even among the Fore—because such cannibalism hasn't been witnessed by "trained observers." But there are many behaviors that humans carry out that are never witnessed by the so-called experts because custom dictates that they be carried out in private—as Jared Diamond pointed out, if he had to rely on observation alone, he would have to conclude that people in New Guinea never engage in sexual intercourse, as despite many years of living in New Guinea while carrying out fieldwork, he never saw such behavior (Diamond 2000). Moreover, there is growing archaeological evidence for cannibalism, such as cut marks on bones that indicate defleshing, polish marks on bones indicating that they had been cooked, and the presence of large amounts of human myoglobin (a protein found in muscle tissue—i.e., flesh) in fossilized human

feces (Marlar et al. 2000). But regardless of whether or not cannibalism was more prevalent in the past, there is certainly no evidence from the prion protein gene that repeated epidemics of kuru-like neurodegenerative diseases plagued our ancestors.

In this section, we have focused on Tajima's D, but in fact there are several such tests of neutrality that rely on different properties of DNA sequence data. For example, Fu and Li's D statistic is based on comparing the number of mutations on the internal branches of a phylogenetic tree versus those that occur on the external branches (tips) of the tree to the pattern expected under neutrality; Fay and Wu's H statistic detects derived alleles that are present at higher frequency than predicted by neutrality (and hence requires an outgroup or other information to distinguish ancestral from derived alleles). And although we have focused primarily on how violating different assumptions of the Tajima's D test can easily result in misinterpretations of the results, don't be fooled into thinking that this is an issue only with this particular test. All statistical tests make simplifying assumptions that you need to be aware of (and beware of) in order to properly interpret the results. Tajima's D remains a widely used and extremely useful way of testing for neutrality—but if you do come across a study claiming that a significant Tajima's D value shows that selection has operated on a particular gene, be sure to check that some other assumption of the test hasn't been violated instead of the neutrality assumption.

SELECTION TESTS BASED ON COMPARING DIVERGENCE TO POLYMORPHISM

The final methodology we will discuss for detecting species-wide selection compares between-species divergence to within-species polymorphism, which basically combines the two approaches discussed so far (i.e., dN/dS ratios are based on divergence between species; Tajima's D and related tests are based on polymorphism within species). The idea is simple: under neutrality, the amount of divergence in the DNA sequences between species should be proportional to the amount of polymorphism among DNA sequences within species, as both depend on the (neutral) mutation rate. Moreover, it has been shown that tests that take both divergence and polymorphism into account have, in general, the most power to detect selection (Zhai et al. 2009).

There are two widely used tests that incorporate both divergence and polymorphism data. The first is the McDonald–Kreitman (MK) test, named after the population geneticists who proposed the test, John McDonald and Marty Kreitman (McDonald and Kreitman 1991). To carry out the MK test for a gene, you construct a 2 × 2 table (Figure 17.5) in which the entries consist of the number of synonymous and nonsynonymous substitutions between species for your gene of interest, and the number of synonymous and nonsynonymous polymorphisms within species. An

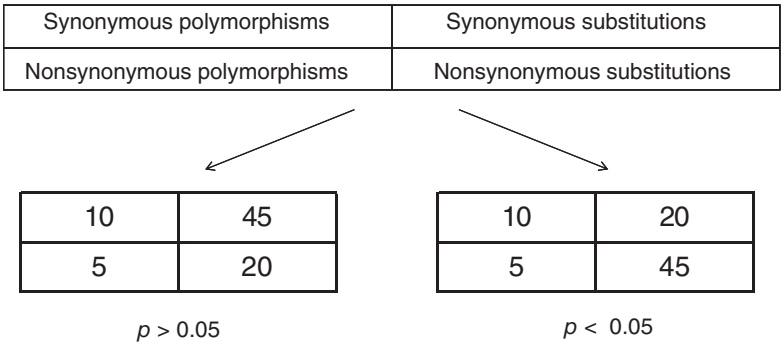


FIGURE 17.5

Logic behind the McDonald–Kreitman test of neutrality. The test consists of a 2 × 2 table, with the entries consisting of the number of synonymous polymorphisms and nonsynonymous polymorphisms observed for a gene within a species, and the number of synonymous substitutions and nonsynonymous substitutions (i.e., fixed differences) between the species of interest (e.g., humans) and a closely related outgroup (e.g., chimpanzees). This is then treated as a 2 × 2 contingency table, for which the probability that the ratio of synonymous versus nonsynonymous polymorphisms is the same as the ratio of synonymous versus nonsynonymous substitutions can be calculated. The null hypothesis is that these two ratios are the same (which is the neutral expectation); rejecting the null hypothesis, therefore, rejects the hypothesis that the gene has been evolving neutrally. Two examples are shown, with the left table consistent with neutrality, but the right table rejects the hypothesis of neutrality due to an excess of nonsynonymous substitutions relative to nonsynonymous polymorphisms, as would be expected if positive selection for nonsynonymous substitutions had been influencing this gene.

excess of nonsynonymous substitutions, compared to nonsynonymous polymorphisms, would be an indication of positive selection acting on amino acid substitutions. The MK test can also reveal other types of selection—for example, an excess of nonsynonymous polymorphisms, compared to nonsynonymous substitutions, would be consistent with a recent relaxation of functional constraints on the gene. That is, negative (purifying) selection in the past has kept the number of amino acid substitutions relatively low, but circumstances have changed recently such that there is less negative selection on the gene, so more amino acid changes can be tolerated. However, this relaxation of negative selection has been recent enough that the amino acid changes have not become fixed in the species, so they instead show up as an excess of nonsynonymous polymorphisms. It should perhaps come as no surprise that many genes in humans do show an excess of nonsynonymous polymorphisms relative to nonsynonymous substitutions. This probably reflects both some relaxation of functional constraints due to culture (i.e., because of cultural developments such as agriculture or modern medicine, some mutations that were deleterious under prehistoric conditions are no longer selected against) and the recent large increases in population size (which basically means that populations are not in equilibrium, which is one of the assumptions of the MK test—population increase basically means that mutations are occurring faster than they can be removed by selection).

While the MK test is an extremely useful test, it is limited in application to protein-coding genes, because it compares nonsynonymous to synonymous substitutions/polymorphisms. What if we want to detect selection on other regions of the genome? A precursor to the MK test, called the HKA test (for Hudson–Kreitman–Aguade, the researchers who developed the test) allows one to do just that (Hudson et al. 1987). As originally formulated, in the HKA test you compare the ratio of the between-species divergence to the within-species polymorphism for your genomic region of interest to the same ratio for a region of the genome that is known to be evolving under neutrality. This test can then be applied to any genomic region of interest. Genomic regions with high levels of interspecies divergence relative to intraspecies polymorphism are candidates for positive selection, while regions of the genome with low levels of interspecies divergence relative to intraspecies polymorphism are candidates for balancing selection. Based on simulated data, the HKA test has been shown to generally be more effective at detecting selection than the other tests described in this section (Zhai et al. 2009).

However, there is one tiny little issue with HKA tests that may have already occurred to you: how do you go about finding a neutrally evolving genomic region to compare with your genomic region of interest? Researchers have generally used regions that are far from any known protein-coding genes, and that do not seem to be involved in any aspects of gene regulation, but there's the rub: there is a lot we don't know about what noncoding DNA in the genome might be doing, even noncoding DNA that is very far away from any protein-coding gene, so these criteria are not sufficient to guarantee neutrality. Another approach is to simulate data under neutrality—and assuming a demographic history that is appropriate for the species that you want to compare—and use the simulated data in the HKA test. Sounds reasonable—as long as you are confident that the demographic history you are simulating really does correspond to the demographic history of your species, otherwise all bets are off. And that, of course, is something that is not so easy to determine.

With genome-wide data, there is another approach that one can use and that is to generate a genome-wide distribution of HKA values and look for extreme values. This can be accomplished easily enough by dividing the genome into chunks (called "windows") that are big enough to contain enough polymorphic sites to be informative (this can be determined by trying different window sizes) and calculating the HKA value (i.e., ratio of between-species divergence to within-species polymorphism) for each window. Assuming that the majority of the genome is evolving under neutrality, which seems an eminently reasonable assumption, then windows with HKA values that are in the extreme tails of the genome-wide distribution are candidates for selection. The lowest HKA scores should be enriched for genomic regions that have experienced positive selection—in other words, in the genome-wide distribution of HKA scores, we expect to find more genomic regions that have experienced positive selection among the lower HKA scores than among the higher HKA scores. This is because genomic regions under positive selection should show low diversity within a species relative to divergence between species—positive selection should enhance divergence. Conversely, the highest HKA scores should be enriched for genomic regions that have been subject to balancing selection, as these show more diversity within a species than expected, based on divergence between species.

Of course, there are lots of other potential reasons why a genomic region would have unusual HKA scores, so it is hard to tell if any particular genomic region is an outlier because of selection or because of some other reason. One way of dealing with this issue is to focus not on individual genes but rather on genes in the same biochemical pathway (such as food digestion) or on genes that carry out related functions (such as immunity). The idea is that since selection acts on the phenotype, multiple genes involved in a particular

phenotype under selection may have been influenced by that particular selective force. And even though each individual gene may show only a slight signal of selection, too little to show up as a candidate all on its own, there may nonetheless be an overall enrichment in the tails of the distribution of HKA scores for genes in a particular pathway, or genes with related functions.

To detect such enrichment, you can carry out what is known as a **gene ontology** analysis. A gene ontology is a standardized description of gene products, intended to facilitate comparisons between studies, especially by computers. The idea is that if, for example, one study says that a gene is involved in RNA synthesis while another study says that the same gene is involved in transcription, then even though these are the same thing, searches based on just one of these terms are not going to find all of the relevant studies. An ontology is basically just a standard vocabulary of terms, and a gene ontology classifies and describes gene products by their function, where in the cell they are found, and what biological process(es) they are involved in (e.g., metabolism, immune response, etc.). Conveniently, the scientific community has gotten together and established the Gene Ontology (GO) project as a collaborative effort to define the terms of the ontology, annotate genes in terms of this ontology, and provide a set of tools that can be used to carry out analyses based on the GO (The Gene Ontology Consortium 2008).

One analysis that can then be done is to see whether any particular GO categories are enriched in the tails of the distribution of HKA scores (or in the tails of any other genome-wide distribution of a test for departure from neutrality). This type of analysis is not as straightforward as it sounds, as it involves carrying out lots and lots of statistical tests, so there will always be some significant results just by chance (remember, a 5% significance level means that you expect about one significant result just by chance for every 20 tests that you do).

Let's look at an example of a GO analysis of genome-wide data. As part of the Great Ape Genome Diversity Project (Prado-Martinez et al. 2013), population genomicist Aida Andrés and colleagues analyzed complete genome sequences from several humans, chimpanzees, bonobos, gorillas, and orangutans (this study is still unpublished as I write this, but hopefully not as you read this). They carried out several different tests (including MK, HKA, and Fay and Wu's H, among others) by dividing the genome up into chunks of appropriate size for each statistic (or, in the case of the MK test, focussing on the genes) and calculating the statistic for each chunk/gene, thereby generating a genome-wide distribution of the values for each test statistic. They then performed a GO analysis of the tails (i.e., the most extreme upper and/or lower values of the statistic, as appropriate) of each distribution. As it turns out, the GO analysis does come up with some interesting results. Not surprisingly, genes involved in the immune response come up as significantly enriched for signals of both balancing selection and positive selection. This is in keeping with lots of studies that have amply demonstrated that infectious diseases and parasites have bedeviled not only humans but also apes (and indeed, practically all living creatures) throughout the course of their evolutionary history, and coming up with strategies and adaptations to fend off such diseases has obviously kept our genome quite busy. More intriguingly, the GO analysis identified several categories related to neurological development and function that were enriched for genes exhibiting signatures of positive selection. Some of the particular genes involved showed similar signals across all or some of the ape species (including humans as apes, which we are), while others were specific to one lineage; these results suggest that selection on brain development and function has been ongoing throughout ape evolutionary history. Another significant result—only in humans—involves the category "starch and sucrose metabolism," that is, digestion of starch and sugar. This result may reflect positive selection during human evolution for adaptations related to changes in the human diet, especially the starch-rich diet of certain European (think of potatoes) and Asian (think of rice) populations. However, it is easy to get carried away with facile interpretations of significant GO results that are entirely speculative—it is important to keep in mind that while we can come up with plausible stories, we don't really know for sure why genes involved in brain function or starch/sugar metabolism should be enriched for signals of positive selection. Moreover, many results are not so easy to interpret—for example, when looking at genes exhibiting signals of positive selection in three or more ape lineages (in order to analyze potential ape-wide targets of positive selection), the category "structural molecule activity" was the only GO category identified by one of the tests as significantly enriched for such genes. "Structural molecule activity" is defined as "the activity of a molecule that contributes to the structural integrity of a complex or assembly within or outside a cell," for example, the proteins that make up the ribosome or that contribute to the cell wall would fall into this category. Why there should have been positive selection on such genes during ape evolution is not at all clear. Nonetheless, these significant GO results provide a starting point for further investigations.

So, GO analysis is a useful way of gaining insights into particular pathways or processes that may have been influenced by selection. Moreover, GO analysis is not restricted to HKA analyses—GO enrichment can

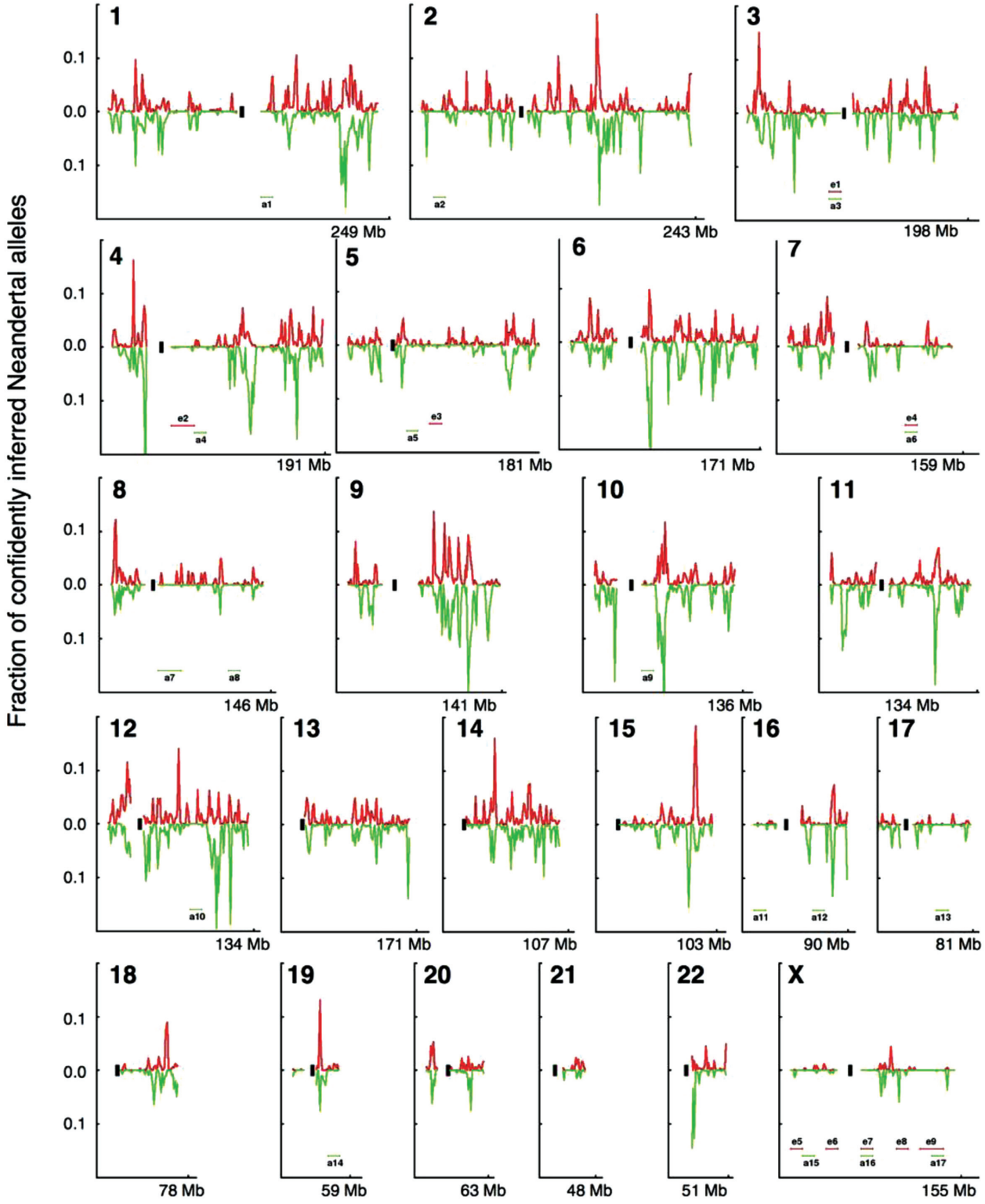


FIGURE 17.6

The landscape of Neandertal ancestry in contemporary Europeans and Asians. The red and green lines show the fraction of alleles confidently assigned (with a probability of at least 90%) as being of Neandertal ancestry in nonoverlapping windows of 1 million base-pairs (1 mB) for each chromosome in Europeans and East Asians (from the 1000 Genomes Project), respectively. The black bar for each chromosome denotes the position of the centromere, and the colored bars indicate 10-mB regions that are deficient in Neandertal ancestry in Europeans (e, red) or East Asians (a, green). Reprinted with permission from Sankararaman, S., et al., "The genomic landscape of Neanderthal ancestry in present-day humans," *Nature* 507:354, 2014.

be applied in any analysis that produces a distribution of scores (dN/dS ratios, Tajima's D values, etc.) in which signals of selection are expected to be concentrated in one part of the distribution (usually, the tails). However, the astute reader may recall that one of the virtues I extolled earlier of generating a genome-wide distribution of HKA scores is that one is not then restricted to investigating only protein-coding genes for signals of selection (as with dN/dS ratios or MK tests, for example) but can also look for such signals on noncoding genomic regions. In principle, this could be even more interesting than selection on genes, as selection on noncoding regions would presumably reflect adaptations involving changes in gene regulation, which (as discussed previously) may be more important evolutionarily than changes in genes. Gene Ontology analysis, however, is by definition restricted to genes. What is needed is an ontology for different types of regulatory elements that can influence gene expression, and while such an ontology alas does not yet exist, great progress is being made in this direction. A large, international consortium with the clever acronym ENCODE (for Encyclopedia Of DNA Elements) is busy cataloging all of the functional elements present in the human genome and developing an ontology that will eventually do for such elements what GO does for genes. A lot of information has already been made publicly available (ENCODE Project Consortium 2012), with more to come, so stay tuned for further developments.

ARCHAIC GENOMES

One final approach to discovering the genetic changes that made us what we are comes from comparisons with archaic genomes. Recently, methods have been developed to identify the actual blocks of archaic ancestry in modern human genomes (Sankararaman et al. 2014; Vernot and Akey 2014; Vernot et al. 2016), and thereby plot the landscape of archaic ancestry across our genomes (Figure 17.6). This analysis reveals regions of the genome with significantly less archaic ancestry than expected; these are referred to as deserts of archaic ancestry and potentially mark regions of the genome where the archaic version was selected against. In other words, deserts provide clues to important genetic differences between archaic and modern humans. As you might guess, figuring out what these might be is an area of active research; intriguingly, the *FOXP2* gene discussed in detail previously falls in one such desert (Vernot and Akey 2014), suggesting that there is indeed an important difference between the archaic and modern human versions of this gene.

CHAPTER 18

LOCAL SELECTION

In the previous chapter, we considered selection that occurred prior to (or during) the origin of modern humans and hence reflects genetic adaptations that are shared by all modern humans. In this chapter, we turn our attention to local selection, meaning genetic adaptations that are specific to a subset of modern humans as they reflect selection due to some local circumstance—environmental, disease-related, diet-related, and so forth. So, what signals can we look for to detect the signature of local selection? In principle, we could use some of the same neutrality tests discussed in the previous chapter, such as Tajima's D statistic, and compare the results in different populations to see whether a particular gene shows a significant departure from neutrality (correcting appropriately for demographic history) in some populations but not others. That would then be a signature of local selection, and we could then investigate those populations with a significant result in more detail. However, the standard neutrality tests are not so good at detecting local selection, as they require extraordinarily strong selection to give a significant result; other tests developed specifically for local selection turn out to be much better at detecting it.

Let's start by considering what local selection actually does in terms of genetic variation at the selected locus (Figure 18.1). Assume that a new mutation arises that has a selective advantage in one population but does not have a selective advantage in another population. Positive selection will then cause the mutation to increase in frequency rapidly in the population where it has a selective advantage and quickly approach fixation, resulting in a selective sweep. With genome-wide data for a population that has experienced the selection and one that has not, there are then two signals of local selection that we can look for. First, the positively selected mutation will show a significantly larger than average genetic distance between the two populations. So, one method of identifying candidates for local selection is to calculate F_{ST} values (or some other measure of genetic differentiation) for genome-wide data and look for outliers, that is, genetic markers with F_{ST} values that are too large to be explained simply by neutrality (Figure 18.2).

The second aspect of genome-wide data that is suggestive of local selection has to do with the associations between a selected allele and nearby variants on the same chromosome (i.e., linkage disequilibrium, discussed in Chapter 9). The basic idea (Figure 18.3) is that a newly arisen mutation is by definition completely associated with all other alleles on the same haplotype. For a neutral mutation, increases in frequency will occur relatively slowly by genetic drift, giving recombination and new mutations ample time to break down the haplotype association (as also shown before in Figure 9.9). As the frequency of a neutral mutation increases, the length of the haplotype associated with that mutation will correspondingly decrease; under neutrality, we expect high-frequency mutations to have short haplotypes. But if a new mutation has a selective advantage, selection will increase the frequency of the mutation more rapidly than under neutrality, thereby decreasing the time for recombination and new mutations to break down the haplotype association. So, a high-frequency mutation associated with a long haplotype is another signature of selection.

Various statistics have been proposed to measure the length of the haplotype associated with an allele. They are all based on the concept of **extended haplotype homozygosity** (EHH), which is defined as the probability that two haplotypes that carry the mutation of interest (usually called the core region) are identical for all other alleles within a specified genomic region surrounding the core (Figure 18.4). There are various kinds of comparisons that can be done that are based on EHH. For example, as shown in Figure 18.4, you can compare the EHH for one allele at a locus with the EHH for the other allele at that locus (assuming a bi-allelic locus, which is usually the case for SNP data). A commonly used statistic that does this is called iHS (which

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

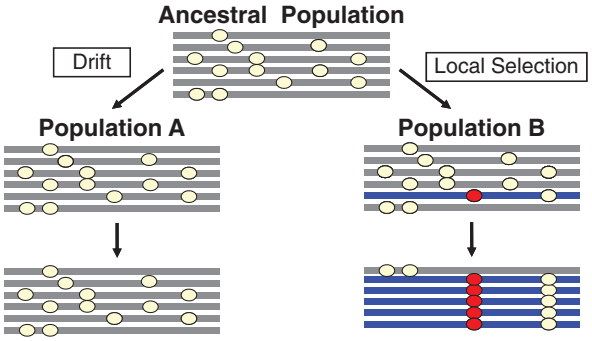


FIGURE 18.1

Effect of local selection versus neutral genetic drift on patterns of genetic variation at a locus. Gray bars are chromosomes and white ovals are mutations. Assume that an ancestral population gives rise to two populations, one of which (A) remains in the ancestral homeland, while the other (B) migrates and experiences different, novel selective pressures (e.g., new diseases). In A, genetic variation is influenced only by genetic drift, but in B a new mutation arises (red oval) that is selectively advantageous and hence increases rapidly in frequency, along with any other mutations on the same haplotype (blue bar). Genetic differences between populations A and B will be much bigger at this locus than for other loci that have not been influenced by local selection.

stands for integrated haplotype score). To compute iHS, you take the EHH curve for one allele at the core SNP and compute the area under the curve that is bounded by some specified value for the EHH (usually 0.05, corresponding to a 5% probability that haplotypes with the allele of interest are identical). Incidentally, computing the area under a curve involves a procedure in calculus known as integration, which is where the "integrated" part of iHS comes from—it doesn't have anything to do with social integration, in case you were wondering. You then divide this by the area under the corresponding EHH curve for the other allele at the core SNP and take the natural logarithm (ln) of this ratio (natural logarithms are logarithms to the base *e*, and *e* was discussed back in Chapter 11). With genome-wide data, you can calculate the iHS score for every SNP and then look for outliers (Figure 18.5), which are then your candidates for local selection.

While iHS is a very useful approach for detecting local selection, it does lack power when the selected allele is at or near fixation, as then the EHH for the nonselected allele cannot be measured very accurately (or at all, if the selected allele is fixed). A complementary approach that is more powerful at detecting sweeps that are at or near fixation is based on comparing the EHH for the same allele in two different populations. The approach is exactly analogous to iHS, except that the EHH curve is plotted for the same allele at the core SNP in different populations (Figure 18.4). The idea is that if an allele is influenced only by neutrality in two different populations, then the EHH for that allele should be more or less the same in those two populations. But if the allele is influenced by neutrality in one population and by positive selection in the other population, then the EHH for the allele in the selected population will be substantially greater than the EHH for the allele in the nonselected population. Various statistics analogous to iHS, such as ln(Rsb) and XP-EHH, have been proposed to detect differences in EHH for the same allele in different populations (and in case you were wondering, it does seem to be mandatory to devise a new and obscure acronym every time you come up with a new statistic!). As with iHS, the idea is that a value for the statistic of choice is then calculated for every SNP in the data set; outliers in the genome-wide distribution are then potential candidates for local selection (Figure 18.6).

While several other methods exist for detecting local selection, they generally rely on the signals discussed previously, for example, extreme population differentiation and/or long EHH. There are also approaches that combine several different measures of local selection into a single "composite" score, which probably provides the most powerful method for detecting local selection (Grossman et al. 2010). Numerous studies have appeared in the past 10 years or so that apply a particular method to genome-wide data (usually SNP chip data) from a set of populations and then come up with a set of candidate genomic regions for local selection because they are outliers—that is, they exhibit values in the distribution(s) of the test statistic(s) that are larger than some arbitrarily chosen cutoff value. For example, you might calculate F_{ST} values for every SNP in your genome-wide data set and then choose the SNPs with F_{ST} values in the top 1% of the distribution as your candidates for local selection.

However, these "outlier" studies have two important limitations. First, it is not clear how to distinguish those genomic regions that actually have been influenced by local selection from false positives due to extreme genetic drift or other demographic effects. We have already seen in the previous chapter that increases in population size can result in significant values for Tajima's D statistic that could be misinterpreted as recent positive selection. Moreover, population geneticist Laurent Excoffier and colleagues have shown that range expansions (population expansions in space as well as in time) can give rise to a phenomenon called "allele surfing" (Klopfstein et al. 2006), in which an allele can quickly rise to high frequency in a small population that is on the frontier of the expansion, giving rise to correspondingly

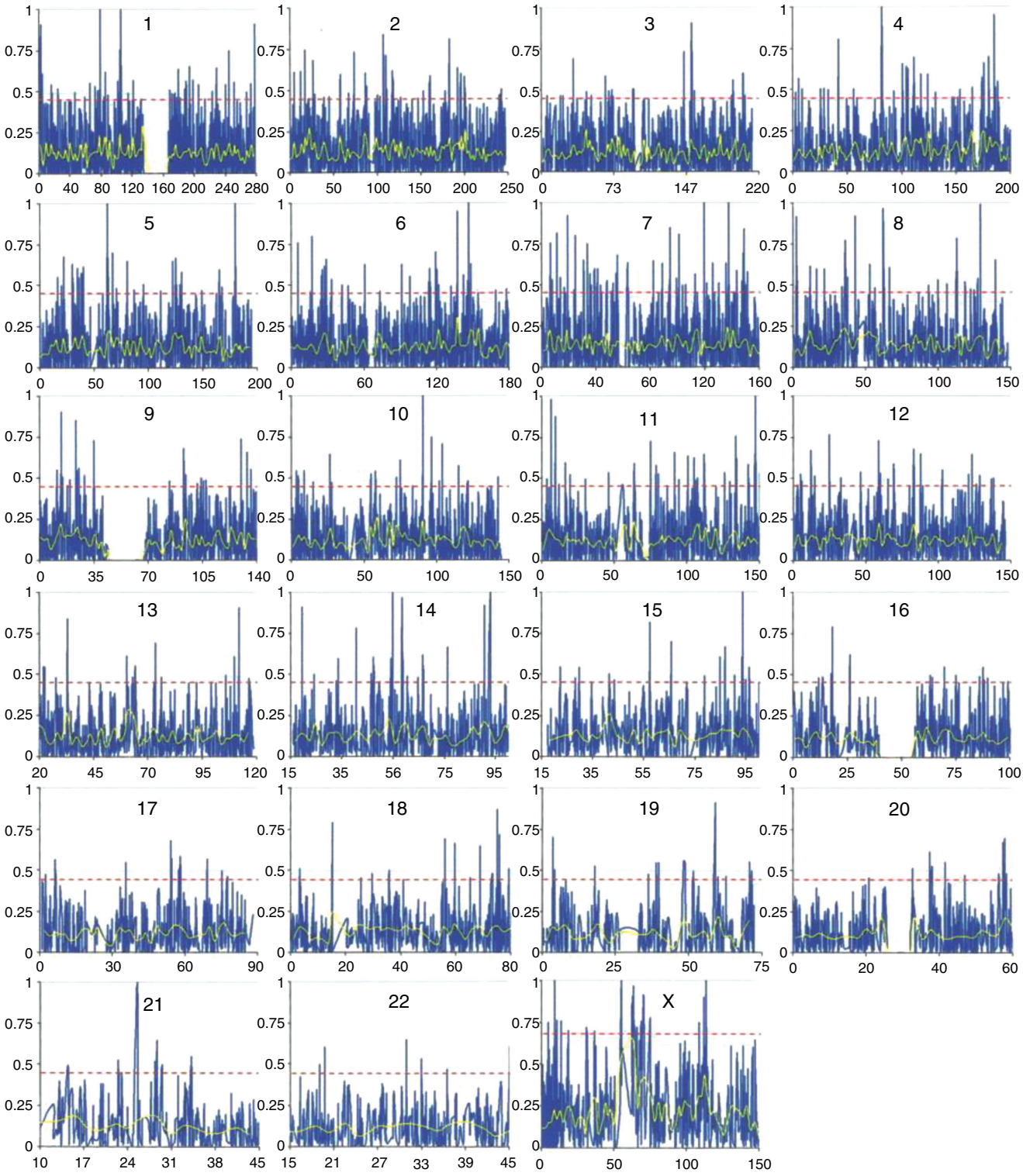


FIGURE 18.2

Example of genome scan for extreme F_{ST} values. Each panel is a chromosome and each blue line is an SNP, with the chromosomal position of the SNP on the X-axis, and the F_{ST} value on the Y-axis. The red dotted line shows the cutoff value for presumed significance (upper 2.5% of the empirical distribution), and the yellow line shows the average F_{ST} value over windows of 1 million base-pairs. Gaps reflect the paucity of data in highly repetitive regions near centromeres. Note the higher average F_{ST} values (and correspondingly higher empirical significance level) for the X chromosome, which presumably reflects either the smaller effective population size for the X chromosome, more opportunities for selection on the X chromosome, or both. Reprinted with permission from Akey, J., et al., "Interrogating a high-density SNP map for signatures of natural selection," *Genome Research* 12:1805, 2002.

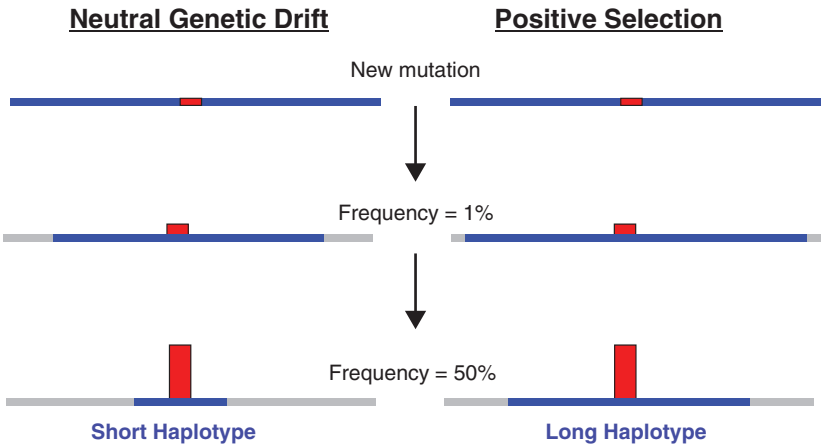


FIGURE 18.3

Haplotype length expected under neutral genetic drift versus positive selection. The red box is the position and frequency of a mutation, the blue bar is the length of the associated haplotype, and the blue + grey bar is the entire chromosome. Under drift, as a mutation increases in frequency, the associated haplotype becomes shorter due to recombination and new mutations, whereas with positive selection, the mutation increases in frequency more rapidly, and hence has a longer associated haplotype.

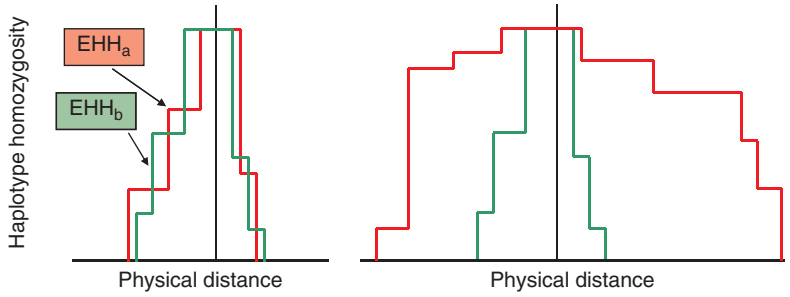


FIGURE 18.4

Haplotype homozygosity plots for a neutral (left) versus positively selected (right) SNP. EHH is the extended haplotype homozygosity plotted from a core SNP (vertical black line), where a and b can refer either to the EHH associated with two alleles at a SNP in a single population (i.e., alleles a and b) or to the EHH associated with the same allele in two different populations (i.e., populations a and b). In the example with positive selection (right), EHHa is greater than EHHb, suggesting positive selection on allele/population a relative to allele/population b. EHH indicates extended haplotype homozygosity.

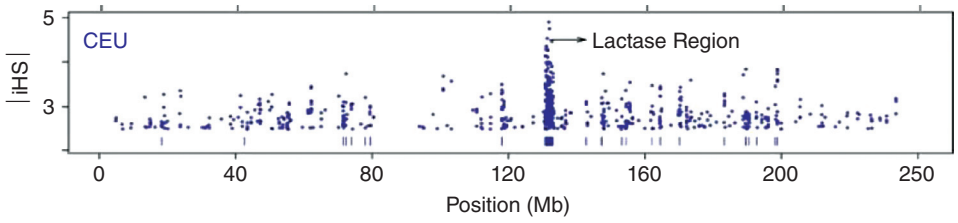


FIGURE 18.5

iHS scores for SNPs on chromosome 2 in a European population. Each dot is a SNP, with the X-axis showing the physical position of the SNP on chromosome 2, and the Y-axis showing the iHS score—this corresponds to the situation in Figure 18.4 in which a and b refer to the two alleles at each SNP within a single population. The vertical lines indicate outliers (in the top 1% of the genome-wide distribution); note that many outliers occur in the lactase region, which will be discussed later in this chapter. CEU indicates Central Europeans; iHS, integrated haplotype score. Reprinted with permission from Voight, B.F., et al., "A map of recent positive selection in the human genome," *PLoS Biology* 4:e72, 2006.

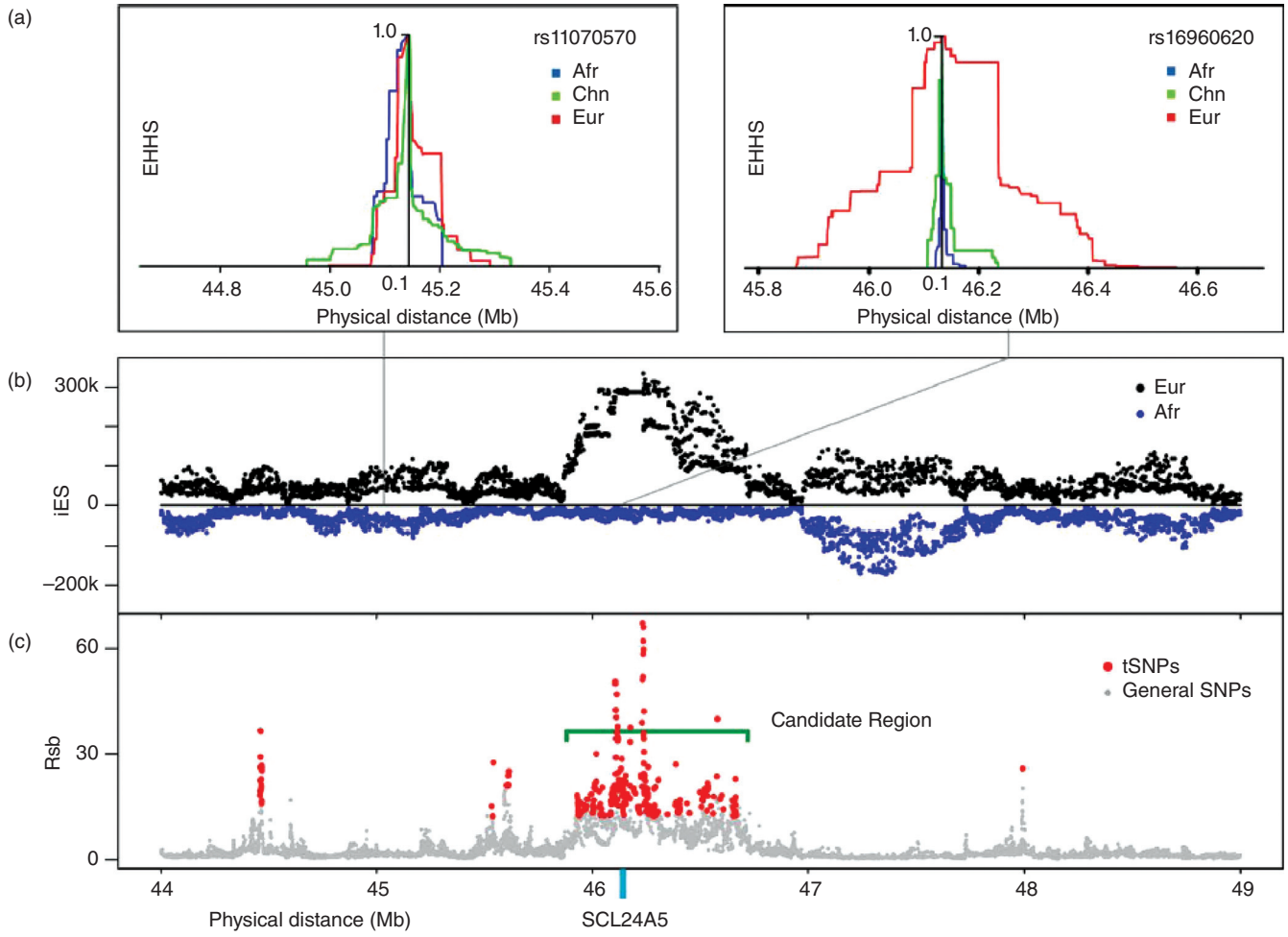


FIGURE 18.6

Example of an extended haplotype homozygosity (EHH) analysis for the same allele in different populations. (a) The left panel shows the EHH curves for the same core SNP allele in three populations (Afr, African; Chn, Chinese; Eur, European); these EHH curves are not significantly different, as expected for a neutral allele. The right panel shows the EHH curves for a SNP allele that shows greatly elevated EHH in Europeans. (b) Each dot is the integrated EHH score (iHS) value for Europeans (black) and Africans (blue), shown as a mirror plot (i.e., European values are plotted above the horizontal axis and African values are plotted below the axis). (c) Values of the Rsb statistic, which is based on the ratio of iHS values in Africans versus Europeans; each dot is the Rsb value for the derived allele at each SNP, with the X-axis indicating the physical position of the SNP on the chromosome and the Y-axis the Rsb value. Red dots are outliers (top 1% of the Rsb distribution) and are concentrated in a candidate region that includes the *SCL24A5* gene (position indicated); *SCL24A5* has been previously implicated in skin pigmentation differences between Europeans and other populations. Reprinted with permission from Tang, K., et al., "A new approach for using genome scans to detect recent positive selection in the human genome," *PLoS Biology* 5:e171, 2007.

high F_{ST} values that might be misinterpreted as local selection.

In fact, it is possible that none of the outliers identified by a genome scan approach have been influenced by local selection, as every distribution will always have a top 1% (or whatever cutoff value is used). That is, we could simulate genetic data under neutrality (without any selection) and generate a distribution of F_{ST} values and take the top 1% as our candidates, but none of these would have actually been influenced by local selection. One way around this problem is to simulate genetic data that are similar to the data to be analyzed (e.g., genome-wide SNP data) under the appropriate demographic model for the populations analyzed and then identify the outliers that are candidates for local selection by comparing the empirical distribution for the statistic (e.g., F_{ST} values) to the simulated neutral distribution for the same statistic.

Sounds reasonable in theory, but in practice it isn't so straightforward (as the baseball philosopher Yogi Berra said, "In theory there is no difference between theory and practice. In practice there is."). What is the appropriate demographic model to use? This often isn't known, at least not with any degree of certainty. As we saw back in Chapter 12, we can use genomic data to estimate the demographic history, so we could use our genome-wide data to first come up with a model of demographic history and then simulate data under that model to get the neutral distribution of our statistic of interest. But if we use genome-wide data that consist of some SNPs that have evolved under neutrality and some that have evolved under selection to estimate a demographic history that assumes that all of the data evolved under neutrality, we will end up with a model of demographic history that is biased against detecting outliers due to selection (because those same outliers will have been used to develop a demographic model assuming neutrality). At present, there is no satisfactory resolution of the debate between those advocating outlier approaches based solely on empirical distributions and those advocating incorporating demographic models. The pragmatic approach that most investigators have adopted is that if there indeed has been local selection, then for sure the best candidates are those with the most extreme values of the test statistic, so let's start with those. Just keep in mind that not all outliers are due to local selection, and that everything that is not an outlier is not necessarily neutral.

The second limitation of genome scans for outliers is that they typically produce a laundry list of candidates for local selection, but it is not at all clear how one should then proceed. Which candidates are worth following up on? Often, genome scans identify relatively large candidate genomic regions encompassing millions of base-pairs that contain numerous genes and/or functional elements; how can the mutation responsible for the signal of selection be identified? Given that a candidate mutation can be identified, what is the functional effect of the mutation, and what is the associated phenotypic effect? And most importantly, *why* has selection acted on this mutation? This is a daunting task for which there are currently no good answers, but before we get too depressed, there have been some success stories, so let's go through two examples where we have been able to learn quite a bit.

EXAMPLE: LACTASE PERSISTENCE

The "textbook" example of local selection in humans is lactase persistence, and so since this is a textbook, we will also go through this example. The story goes as follows: lactose is the major sugar that is present in mammalian milk and is broken down in the small intestine by the enzyme lactase into two other sugars, glucose and galactose. These are then absorbed into the bloodstream and the galactose is subsequently converted to glucose, which is the primary source of energy for the body. The usual situation in mammals is for lactase to be highly expressed only while mother's milk is the primary source of nutrition; after weaning, levels of lactase decrease significantly. This is a classic example of the evolutionary principle of "use it or lose it": since most mammals no longer encounter milk (or lactose) as a nutritional source after weaning, there is no benefit to the continued production of lactase after weaning. However, some of us humans are weird in that we retain high levels of lactase expression into adulthood, long after we are weaned from our mother's breast, and this trait is called **lactase persistence** (LP) or **lactose tolerance**. Individuals who are LP are able to digest milk without any problems, whereas individuals who do not retain high levels of lactase expression into adulthood are said to be **lactose-intolerant** and frequently (but not always) suffer from stomach upset, abdominal cramps, diarrhea, and/or flatulence after drinking milk.

It should therefore come as no surprise that the frequency of LP is associated quite strongly with populations with a history of pastoralism and drinking milk. In particular, LP is highest in frequency in European populations (Figure 18.7) and also in certain populations in Africa (more on this later). Across Europe, LP increases in frequency from southeast to northwest, which is also in accord with a higher reliance on milk drinking as a source of nutrition in northern Europe than in southern Europe.

The genetics of LP was worked out in the 1970s, and LP was shown to be inherited as an autosomal dominant trait (reviewed in Ingram et al. 2009). Further biochemical studies during the 1980s verified autosomal dominant inheritance, as a single copy of the LP allele is sufficient for lactose tolerance (reviewed in Ingram et al. 2009). Since LP is a simple Mendelian trait, it was expected that there would be a single mutation responsible for LP, and hence DNA sequencing of the lactase gene should readily reveal this mutation. It therefore came as something of a surprise when DNA sequencing studies in the 1990s did not find any mutations in the lactase gene that were associated with LP (reviewed in Ingram et al. 2009). The mystery was cleared up in 2002 when more extensive sequencing did find a putative LP mutation (Enattah et al. 2002)—but not in the lactase gene itself! Instead, the LP mutation was found well outside the lactase gene in an intron of a completely different gene (Figure 18.8). This mutation is a C to T substitution at a position that is 13,910 bases upstream of the transcription initiation site for lactase and so is imaginatively known as the *-13910*T* allele. Further studies showed that the

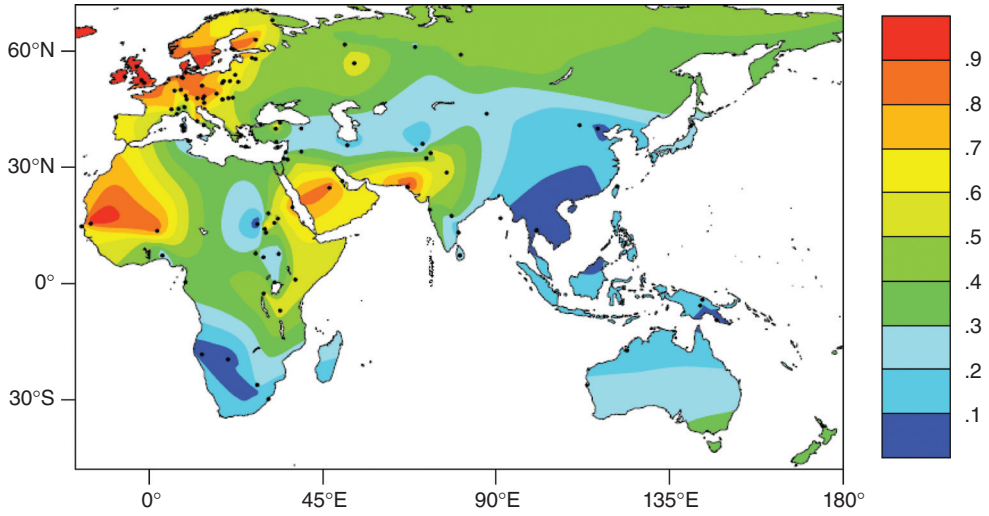


FIGURE 18.7

Frequency of lactase persistence in worldwide populations. The dots show the location of sampled populations, and the colors are interpolated frequencies according to the color bar on the right. Reprinted with permission from Itan, Y., et al., "A worldwide correlation of lactase persistence phenotype and genotypes," *BMC Evolutionary Biology* 10:36, 2010.

*-13910*T* mutation results in increased production of lactase mRNA in cell lines and hence increased levels of lactase expression. Hopefully, it has not escaped your attention that this is exactly the kind of mutation that was mentioned back in Chapter 6, when we discussed the potential evolutionary importance of changes in gene expression. Namely, *-13910*T* is an example of a regulatory mutation that influences how much of a protein is made, rather than a mutation that changes the structure of a protein. So, the lactase persistence

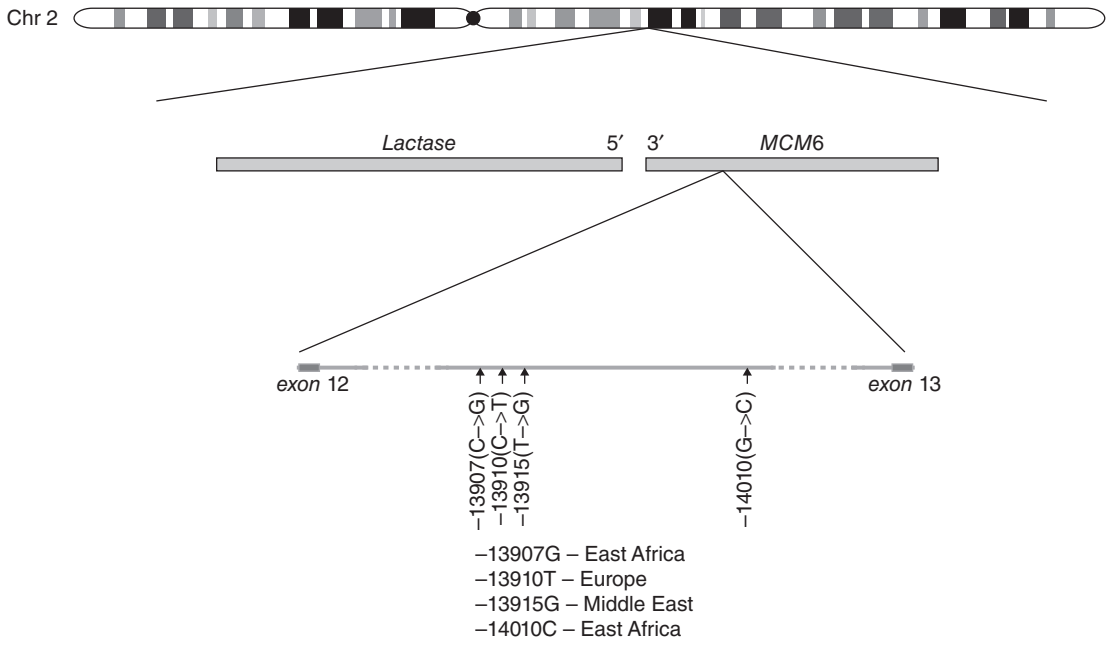


FIGURE 18.8

Chromosome 2, showing the location of the lactase gene and the neighboring *MCM6* gene. Several mutations associated with LP are located in an intron of the *MCM6* gene, about 14,000 bp away from the start of the lactase gene itself. Reprinted with permission from McIntosh, S.K., and Scheinfeldt, L.B., "It's getting better all the time: Comparative perspectives from Oceania and West Africa on genetic analysis and archaeology," *African Archaeological Review* 29:131, 2012.

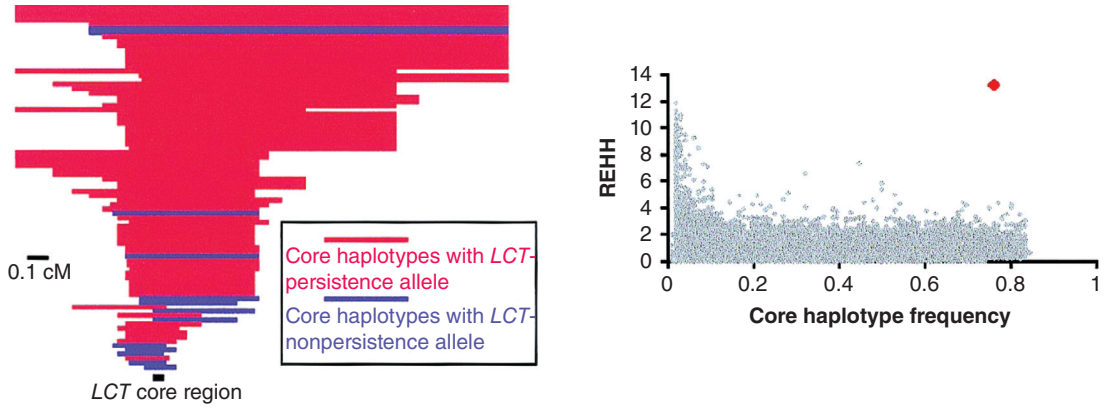


FIGURE 18.9

Left, the relative lengths of haplotypes associated with the lactase persistence (LP) allele (red) and with the non-LP allele (blue) at position -13910 in Europeans. The average haplotype length associated with the LP allele is much longer than that associated with the non-LP allele, even though the LP allele is at much higher frequency. This is the classic signature of recent positive selection. Right, results of simulations of the extended haplotype homozygosity (EHH) (here called REHH) under neutrality for a given core haplotype frequency. Each gray dot is the result of one of 10,000 simulations; the red diamond is the observed EHH for the LP allele, showing that it is extremely unlikely for the observed EHH to have arisen by neutral genetic drift. Reprinted with permission from Bersaglieri, T., et al., "Genetic signatures of strong recent positive selection at the lactase gene," *American Journal of Human Genetics* 74:1111, 2004.

story is a nice example that supports the evolutionary importance of changes in gene regulation.

And where does selection enter the story? It turns out that there is a strong signal of recent positive selection on the lactase gene (and the surrounding genomic region). The *-13910*T* allele is on a haplotype that is much longer than expected, given the high frequency of this allele in European populations (Figure 18.9). Recall from the discussion earlier in this chapter (and Figure 18.3) that a new mutation initially occurs by definition on just one haplotype; over time, as the mutation increases in frequency, the length of the associated haplotype gets smaller and smaller due to recombination and new mutations. Under neutrality, with genetic drift (random changes in allele frequencies from generation to generation) as the dominant evolutionary force, we expect high-frequency alleles to be associated with short haplotypes, as it takes a long time for alleles to drift to high frequency. Positive selection, on the contrary, drives alleles to high frequency quickly, resulting in long haplotypes. The *-13910*T* allele sits on a very long haplotype, much longer than would be expected under a neutral drift scenario, which is strong evidence indeed for recent positive selection on this allele (Bersaglieri et al. 2004).

The above results for the *-13910*T* allele are based on studies of European populations, where the evidence suggests that it is responsible for most (if not all) of the LP trait. However, LP exists outside Europe, most notably in some African populations that make use of milk drinking (Figure 18.7), and here there are big discrepancies between the frequency of LP and the frequency of the *-13910*T* allele (Table 18.1). This suggests that there must be other LP alleles in African populations. And indeed, in 2007 geneticist Sarah Tishkoff and colleagues published a study that looked for mutations associated with LP in East Africans (Tishkoff et al. 2007). This was not so easy as the field test for LP involved asking people to fast for a minimum of 8 hours before the test and then giving them a good slug of lactose dissolved in water (equivalent to about 1–2 L of cow's milk) to swallow, followed by fingerpricks at regular intervals thereafter to collect blood for glucose measurements. The idea is that people with LP will show a pronounced rise in blood glucose levels as the lactose is metabolized to glucose, while lactose-intolerant people will show no such increase. Given that lactose-intolerant people will also tend to feel nauseated and exhibit the other unpleasant signs of inability to digest lactose mentioned previously (such as diarrhea and flatulence), you can well imagine that this was not the most popular test to administer! Nevertheless, Tishkoff and colleagues were able to identify several novel variants—all in the same intronic region as the *-13910*T* mutation—that were correlated with LP (Figure 18.10). These novel variants were also associated with increased expression of the lactase gene in cell lines and, moreover, also showed signatures of strong recent positive selection. Thus, the LP story also provides a very nice example of **convergent evolution—**namely, different mutations

TABLE 18.1 ■ Frequency of the LP trait and the *-13910*T* allele in various populations[a]

Population	Frequency LP	Frequency *-13910*T*
UK Europeans	0.74	0.78
US Europeans	0.74	0.74
France	0.49	0.52
Italy	0.14	0.18
African–Americans	0.11	0.19
Fulbe	0.29	0.26
Hausa	0.23	0.30
Wolof	0.51	0.09
N. Sudan	0.45	0.07
Nuer	0.22	0.07

LP, lactase persistence.

[a]For the non-African populations and some African populations (in black), the frequency of LP is approximately the same as the frequency of the *-13910*T* allele, suggesting that this allele is probably the only allele involved in LP in these populations (the frequency match is not exact because different individuals from these populations were used to measure the LP trait and the allele frequency). For some African populations (red), the frequency of the LP trait is significantly higher than the frequency of the *-13910*T* allele, suggesting that additional mutations contribute to LP in these populations. *Source:* Swallow, D.M., "Genetics of lactase persistence and lactose intolerance," *Annual Review of Genetics* 37:197, 2003; Mulcare, C.A., et al., "The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C−13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans," *American Journal of Human Genetics* 74:1102, 2004.

giving rise to the same phenotypic trait in different populations.

So, here we have a nearly complete story: mutations have been identified that cause the retention of lactase expression into adulthood (although the exact molecular mechanism has not been worked out, there is some evidence that implicates binding of transcription factors), thereby allowing consumption of cow (and other mammalian) milk without any nasty side effects, and there is a strong signal of recent positive selection for these mutations. There is only one piece missing, but it is a big one, and that is: *why* has there been such strong selection for these LP mutations? The obvious answer, given the association between LP and populations that drink milk, would seem to be that having access to milk as a source of food into adulthood was a significant advantage. Indeed, this has become known as the "culture historical" hypothesis (Simoons 1970), and to be sure there is plenty of evidence in its favor. For example, the age of the *-13910*T* mutation (based on the amount of linked variation) is about 7,000–12,000 years ago (Coehlo et al. 2005), which fits nicely with dates of about 9000 years ago or so for the onset of dairying in Europe (Vigne 2008).

But some researchers have questioned whether milk really provided such a significant food source. Keep in mind that the strong signal of recent positive selection means that people with the *-13910*T* allele were leaving more surviving offspring than people without it. But people who are lactose-intolerant do not seem to be so badly off, as long as they avoid drinking milk, so would they really be leaving fewer surviving offspring? Of course, we don't really know how things were during the early Neolithic, when dairying was just starting. Maybe having year-round access to milk was a significant nutritional advantage for people who were able to drink it. Still, some researchers have proposed alternative explanations for the strong selection on LP alleles (reviewed in Gerbault et al. 2011).

One hypothesis is known as the "arid climate" hypothesis and points to the value of milk as an alternative source of water, not food. According to this hypothesis, LP alleles in Europe may have had their origin in the Middle East and other arid environments, where access to clean water can be a big problem, especially for desert nomads. If you could then drink milk without the lactose-intolerant side effects—especially during outbreaks of diarrhea or other waterborne diseases—then you would be much better off. However, while this hypothesis may explain the strong selection for LP alleles in Africa and other dry places, it cannot account for the selection for LP alleles in Europe—especially northwestern Europe, where LP has the highest frequency—as access to clean water was generally never a problem.

An alternative hypothesis for Europe is that the selection for LP is related to calcium uptake and absorption. Calcium is an essential mineral for healthy bones, and if you don't get enough calcium, you can develop rickets, which is a softening of the bones that can lead to increased fractures. It is particularly important for pregnant mothers to get enough calcium in the diet, as otherwise the developing fetus will actually leach calcium from the mother's bones, which in extreme cases can endanger the mother's life (fetuses can be nasty, selfish creatures!). Vitamin D also plays an important role in calcium uptake, and vitamin D is synthesized by your skin in the presence of sunlight. The low sunlight levels typical of northern Europe are thought to be responsible for decreased vitamin D production and increased susceptibility to rickets (indeed, one popular hypothesis for the lighter skin pigmentation in Europeans is that it facilitates vitamin D synthesis—more on this in Chapter 20). Therefore, according to the "calcium uptake" hypothesis, milk provided an important source of calcium (as well as small amounts of vitamin D) that was not available to non–milk drinkers. However, processed dairy products such as cheese or yoghurt contain reduced

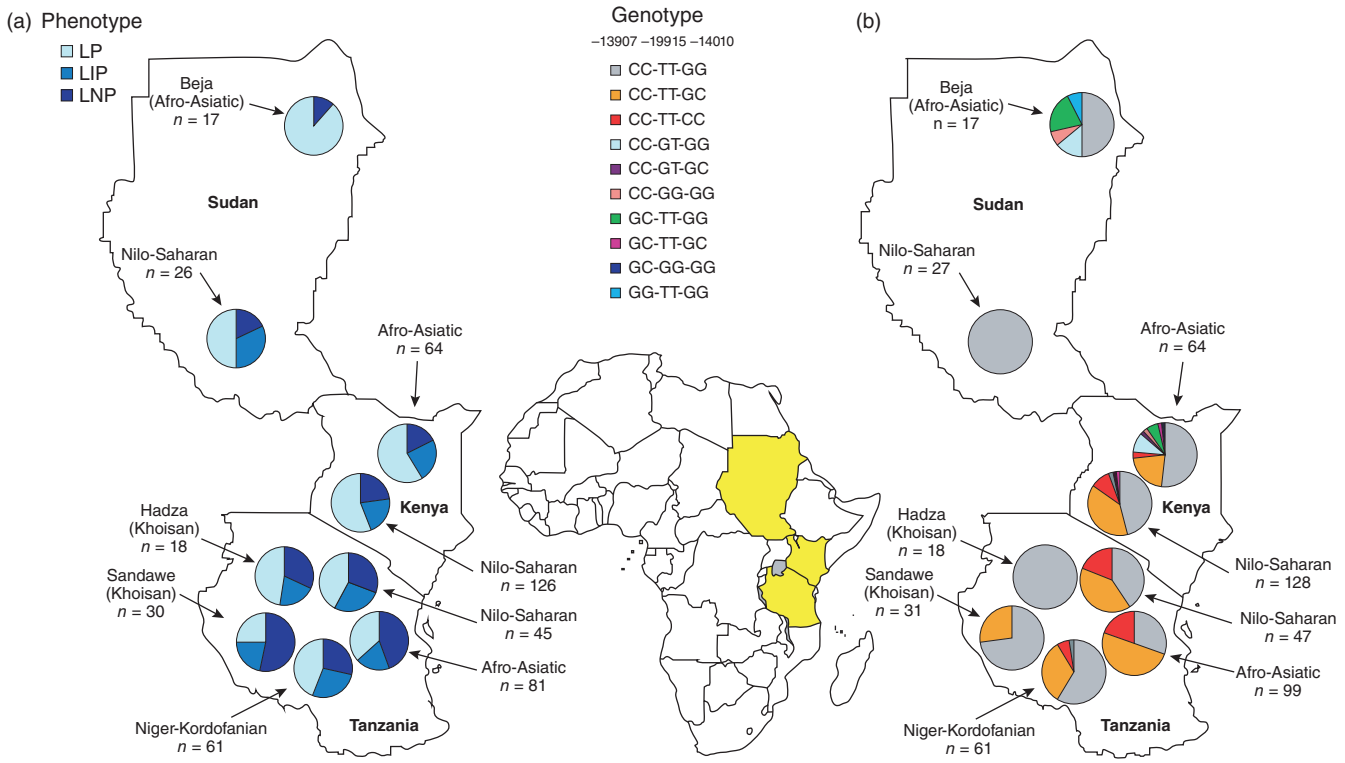


FIGURE 18.10

Association between LP in African populations from Sudan, Kenya, and Tanzania and novel LP variants. The pie charts (a) in the map indicate the frequency of lactase persistence (LP), lactase nonpersistence (LNP), and lactase "intermediate" persistence (LIP, defined as an equivocal response to the test for lactose tolerance, which involves administering 50 g of lactose to the subject and then monitoring the rise in blood glucose levels over time). Center, haplotypes for three novel variants (−14010C, −13915G, and −13910G) that are associated with LP in these populations and elevate levels of lactase expression in cell lines; the pie charts (b) in the map show the haplotype frequencies for each population. Modified with permission from Tishkoff, S.A., et al., "Convergent adaptation of human lactase persistence in Africa and Europe," *Nature Genetics* 39:31, 2006.

amounts of lactose and hence can be consumed by the lactose-intolerant without ill effects, thereby providing at least some calcium. It's not hard to think up other potential explanations for the signal of strong selection on LP alleles. One idea that arose out of discussions with my students one day is that maybe the selection for LP is related to fertility rather than viability. In so-called natural fertility societies (i.e., those without access to birth control), the major constraint on the number of children that a women can have during her reproductive years is the time she spends nursing her children. The usual state of affairs is for the breast-feeding woman to cease ovulation until the child is weaned, whereupon ovulation resumes and she can then become pregnant again. It is not unusual in many societies for women to breast-feed until the child is 3–4 years old (or even older), which therefore limits the number of children a women can have. Suppose that a child who is lactose-tolerant can be switched from nursing to drinking cow's milk at an earlier age, then the mother would resume ovulation and potentially fall pregnant again sooner than if she nursed the child to the usual age. The end result would be that women who switched their children to cow's milk at an earlier age would end up with more children during their reproductive life span, thereby producing selection on the LP allele because of increased fertility. I have no idea whether this scenario could actually produce the strong signal of selection observed for LP, but it certainly sounds good!

Anyway, this multitude of potential explanations for the strong selection for LP illustrates one of the biggest issues with selection studies, namely, the difficulty with trying to answer the "why" question: *why* was there selection for a particular allele or trait? When we study selection, this is usually what we ultimately want to know, but at the same time, it is not at all clear how we can actually test competing explanations (or even come up with an explanation in the first place). We'll revisit this issue at the conclusion of this chapter, after we take a look at another example of local selection.

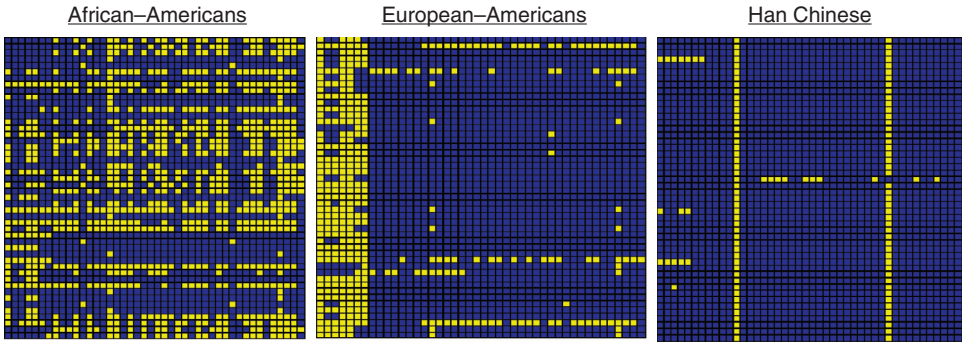


FIGURE 18.11

Visual haplotype graphs for the *EDAR* gene in African–Americans, European–Americans, and Han Chinese. Each row is a haplotype and each column is a SNP, with blue and yellow indicating the allelic state of each bi-allelic SNP. The sequence data used to generate these graphs are from the Seattle SNPs Web site (http://pga.gs.washington.edu/), which was a project to sequence various genes from a standard panel of DNA samples; the project ended in 2009.

EXAMPLE: EDAR

Lactase persistence provides an example of a phenotypic trait of interest for which a genetic basis was then found, followed by indications of strong selection for the underlying mutation(s). However, as discussed previously in this chapter, nowadays scans for selection are all the rage, thereby producing a list of candidate genome regions that have putatively been influenced by recent positive selection, in the absence of any prior knowledge as to what the underlying phenotype might be. So, let's go through an example and see how one proceeds from identifying a candidate region to finding out what is producing the signal of selection—and, maybe, even getting a hint as to why there is a signal of local selection on this candidate region.

One of the most successful stories—in terms of what has been learned—to result from genome scans involves a gene called *EDAR*. EDAR stands for the ectodysplasin A receptor (more on what EDAR does in just a bit) and usually ends up on the list of candidate genes for recent positive selection in pretty much any genome scan that includes East Asian populations (in particular, Han Chinese). No matter what statistic is analyzed (F_{ST}, iHS, ln(Rsb), etc.), you name it, EDAR in East Asians shows up as an outlier. So what is going on with EDAR in East Asians? One way to start investigating patterns of genetic variation at a locus is via a **visual haplotype graph** (VHG). This is, as the name suggests, a convenient way of visualizing the haplotype variation at a locus in a sample of individuals; in such graphs, each row is a haplotype (so, two haplotypes per individual) and each column is a polymorphic site. As can be seen in Figure 18.11, the VHG for African–Americans shows lots of variation (as to be expected, given overall higher levels of genetic diversity in Africans than in other populations, as well as the admixed history of African–Americans), the VHG for Europeans somewhat less variation (in keeping with the reduced genetic diversity in Europeans), but the VHG for Han Chinese shows hardly any variation, which is dramatically different from the usual case for East Asians—most often, they show levels of genetic diversity intermediate between Europeans and Africans. So, it looks like there has been selection for a particular EDAR haplotype in East Asians.

With the DNA sequence information, we can then ask whether there are any mutations of interest that are associated with this haplotype. It turns out that there is one mutation associated with this haplotype that results in an amino acid substitution of alanine for the usual valine at position 370 of the protein (hence the name for this mutation: 370A). The 370A allele is found at high frequency across East Asia and in populations with East Asian ancestry (such as Native Americans) but is virtually absent elsewhere, and the associated F_{ST} values between East Asian and other populations are highly significant (Figure 18.12). This nonsynonymous mutation is thus a prime candidate for somehow being responsible for the signal of selection on EDAR—but how can we learn more?

This is a good point to take further stock of what is known about EDAR. It turns out that mutations have been found which greatly reduce or eliminate EDAR function, and these mutations result in a disease called **ectodermal dysplasia**. There are a variety of different syndromes that fall under this disease, reflecting abnormalities in the development (dysplasia) of features of the exterior layer of the body (the ectoderm). People with ectodermal dysplasia due to EDAR mutations generally exhibit defects in hair structure, sweat

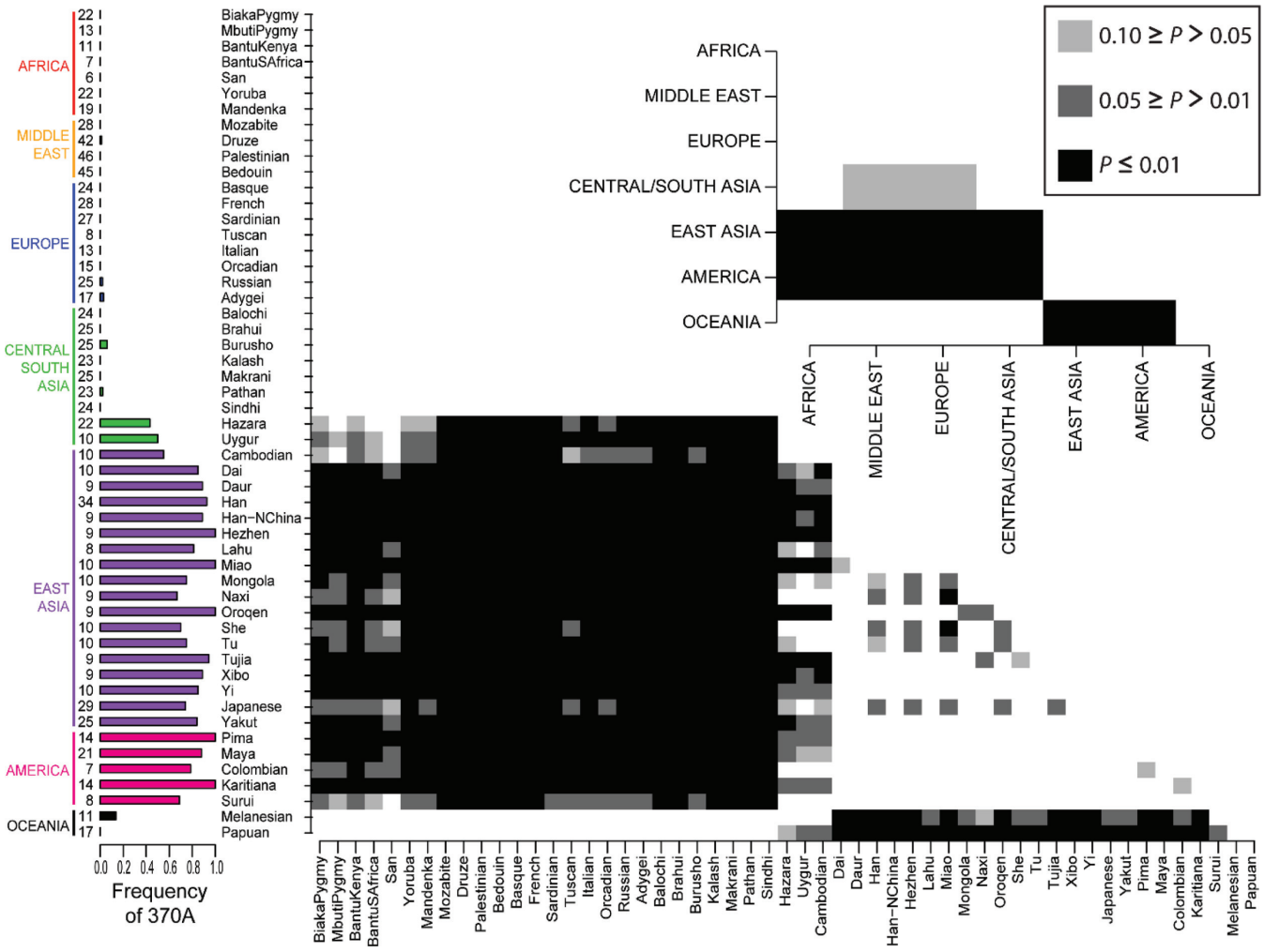


FIGURE 18.12

Frequency of the EDAR 370A allele and associated F_{ST} values in HGDP (Human Genome Diversity Panel) populations. The bars on the left, color-coded according to major geographic region, indicate the frequency of the 370A allele; population names are to the right of the bars and sample sizes are to the left. The heat plot on the right shows the p value (based on the empirical genome-wide distribution of F_{ST} values between each pair of HGDP populations), with the inset showing the corresponding heat plot for F_{ST} values between geographic regions. Reprinted with permission from Bryk, J., et al., "Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation," *PLoS ONE* 3:e2209, 2008.

glands, and teeth. And now this starts to become interesting, because these are all traits in which East Asians do show average differences compared to other populations (especially Europeans). East Asian populations have, on average, thicker hair and fewer apocrine sweat glands than European populations. Apocrine sweat glands are one of two types of sweat glands that humans have (the other are eccrine sweat glands—more on these later), are located only in a few specific parts of the body (such as the armpits), and excrete fat and protein, along with water. They are thought to be associated with body odor and may, therefore, be responsible for the anecdotal perception that East Asians have a less pungent body odor than Europeans (not that I am aware of any study that has actually tried to measure body odor—or, to use the fancy scientific term, "axillary osmidrosis"—in different populations!). With respect to teeth, East Asian populations (and populations with East Asian ancestry, such as native Americans) have higher frequencies of **shovel-shaped incisors** (ridges along the edges of the front teeth that give the back side a shovel or scoop-like shape, as shown in Figure 18.13) than other populations. So, perhaps the 370A mutation has something to do with the differences in these traits between East Asian and other populations?

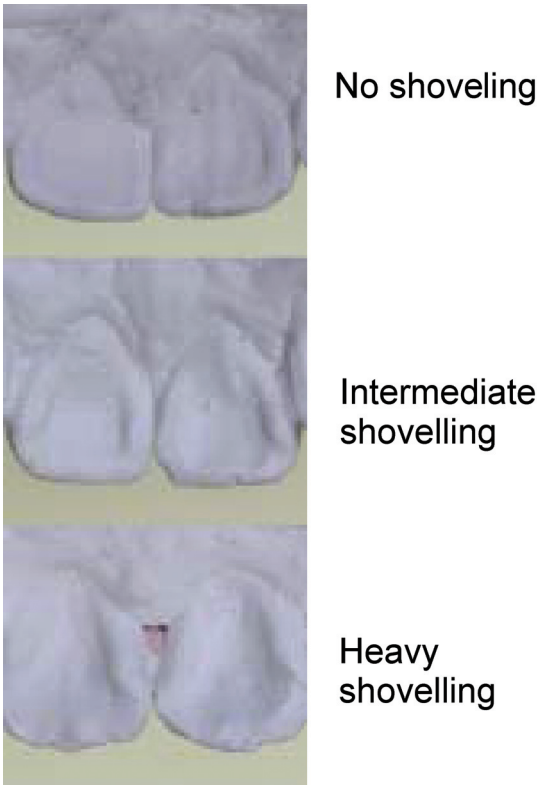


FIGURE 18.13

Variation in shovel-shaped incisors. The degree of shoveling increases from top to bottom. Reprinted with permission from Kimura, R., et al., "A common variation in EDAR is a genetic determinant of shovel-shaped incisors," *American Journal of Human Genetics* 85:528, 2009.

The first indication that this might indeed be the case came not from studying humans but rather from mice. Denis Headon, a researcher who was studying how different ectodermal appendages (hair, feathers, antennae, horns, antlers, etc.) are developed in different creatures, decided to see what would happen in a mouse that was genetically engineered to express higher levels of mouse EDAR (since EDAR was known to be involved in ectodermal development). The result (Mou et al. 2008), shown in Figure 18.14, was a mouse with thicker hair than normal. This study was quickly followed by association studies in humans—that is, studies that compared hair thickness in people with and without the 370A allele—that came to the same conclusion (Fujimoto et al. 2008): the 370A allele is indeed associated with thicker hair (Figure 18.15). Soon after, a study came out reporting that the 370A allele is also associated with shovel-shaped incisors (Kimura et al. 2009).

At the same time as these phenotype association studies were being done, researchers also started looking into the impact of the 370A allele on EDAR

FIGURE 18.14

Effect of elevated expression of endogenous mouse EDAR. (a) Comparison of wild type mouse (left) with the enhanced EDAR mouse (right). (b) Comparison of hair shafts from wild type (left) and enhanced EDAR (right) mice. Reprinted with permission from Mou, C., et al., "Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form," *Human Mutation* 29:1405, 2008.

function. It turns out that EDAR is part of a signaling pathway that is crucial for the development of certain ectodermal structures—in particular, hair, sweat glands, and teeth. This pathway is diagrammed in Figure 18.16, so let's see what happens. Recall that EDAR stands for "ectodysplasin A receptor," so the protein encoded by the *EDAR* gene is a cell-surface receptor—that is, a protein that sits partly inside the cell and partly outside the cell (and part of the protein thus passes through the cell membrane). The part outside the cell binds to a specific ligand, which in this case is the ectodysplasin A (EDA) protein, made by the *EDA* gene. When EDAR binds EDA, it undergoes a change in shape, not just on the outside of the cell but also on the inside of the cell. This, in turn, leads to binding of a protein called EDARADD to the part of EDAR on the inside of the cell. EDARADD stands for "EDAR Associated Death Domain," because the part of EDAR that EDARADD binds to is known as the "death domain," for reasons that will be explained shortly. Once EDARADD binds to EDAR, a signaling cascade is set off inside the cell (basically, proteins bind to other proteins and cause shape changes in those proteins that lead to further protein binding and shape changes, and so on), culminating in the activation of a protein called NF-κB, which is a transcription factor. When the inactive NF-κB gets activated by the EDAR signaling cascade, it moves into the nucleus of the cell and initiates transcription at various target genes. The fact that the ultimate function of EDAR is to activate a transcription factor, which in response turns on the expression of many other genes, provides a potential clue as to how a mutation in a single gene (EDAR) can have so many phenotypic effects (e.g., on hair structure, sweat glands, teeth, etc.).

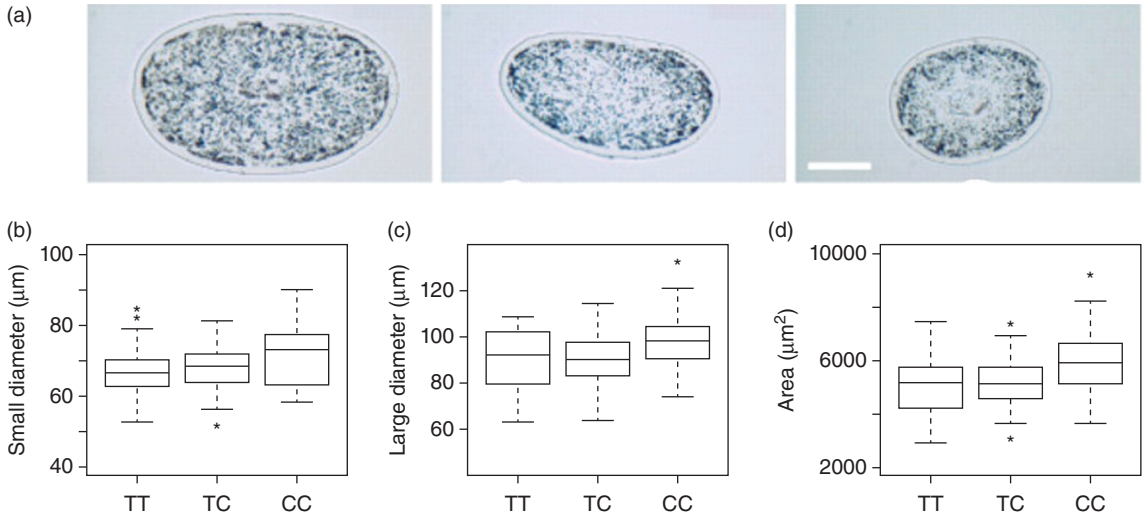


FIGURE 18.15

Hair characteristics associated with EDAR genotypes in humans. (a) Cross sections of typical hairs from individuals homozygous for the 370A allele (left), heterozygous 370A/V (middle), and homozygous for the 370V allele (right). (b–d) Box plots for different hair measurements for the three corresponding EDAR genotypes (TT, TC, and CC, respectively). Note that a box plot depicts the overall distribution of values (in this case, hair measurements for each genotype): the box encompasses the 25–75% range, the median value is depicted by the thick horizontal line, and the extended lines (sometimes referred to as whiskers) show the maximum and minimum values, with outliers shown as individual dots. Reprinted with permission from Fujimoto, A., et al., "A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness," *Human Molecular Genetics* 17:835, 2008.

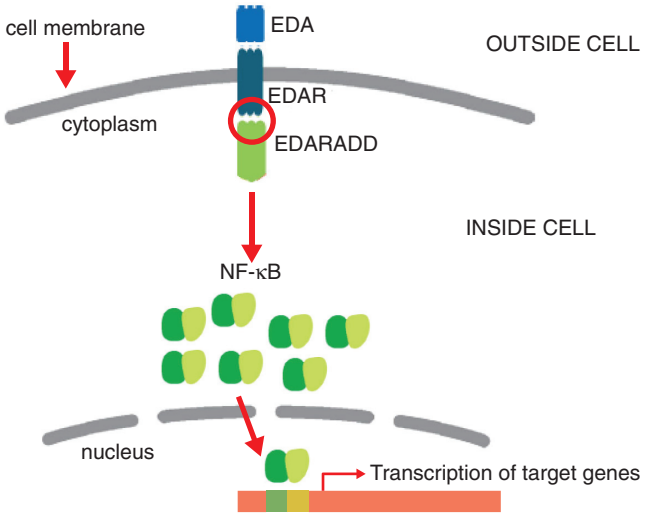


FIGURE 18.16

EDAR signaling pathway. When the EDA ligand binds to EDAR outside the cell, it induces a shape change in the EDAR death domain inside the cell, which leads to binding of the EDARADD protein. This in turn initiates a signaling cascade, with the ultimate outcome that a transcription factor, NF-κB, is activated, moves inside the nucleus of the cell, and initiates transcription of various other target genes.

One of the key steps in the EDAR signaling pathway is thus the binding to EDARADD, after EDAR has been "activated" by binding to the EDA ligand. The part of the EDAR protein that EDARADD binds to is called the death domain because similar domains were first discovered in other proteins, and when they were activated, they caused the cells to die (cells causing their own death also goes by the hard-to-pronounce term **apoptosis**). It may seem strange to you that cells would willingly cause their own death, but in fact regulated cell death is a very important part of growth and development. For example, our separate fingers and toes are formed by the death of cells in between them; defects in this cell death process can lead to webbed fingers or toes. However, keep in mind that even though for historical reasons the part of EDAR that EDARADD binds to is called the death domain, in the case of EDAR the result is the (eventual) activation of a transcription factor, not cell death.

Several lines of evidence indicate that the death domain plays a crucial role in EDAR function. As shown in Figure 18.17, the amino acid sequence of the death domain is highly conserved between species—in fact, there are just two amino acid differences between the human and chicken forms of the death domain, much less than would be expected based on the average difference between human and chicken proteins.

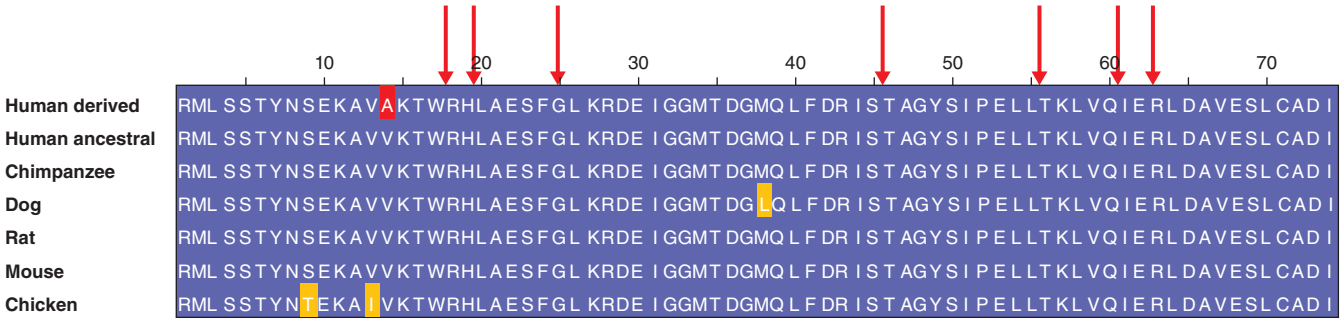


FIGURE 18.17

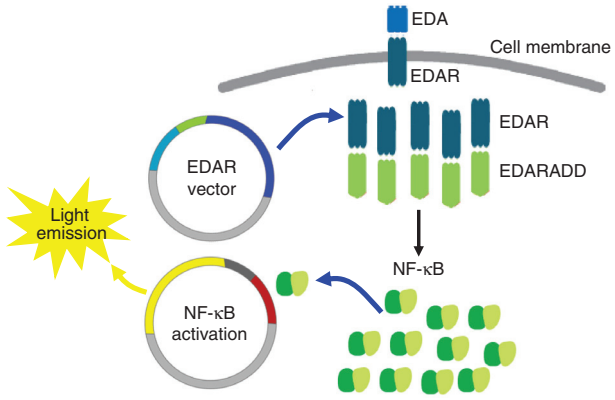
Amino acid sequence of the death domain of EDAR from various species. This domain is highly conserved, with no differences between the human and rat or mouse sequences, and only one difference between the human and dog sequences and two differences between the human and chicken sequences (highlighted by yellow boxes). The arrows indicate the positions of mutations that cause ectodermal dysplasia by reducing or eliminating EDAR function, and the red box indicates the A/V polymorphism at position 370. Reprinted with permission from Bryk, J., et al., "Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation," *PLoS One* 3:e2209, 2008.

Recall from Chapter 6 that conservation is an indication of functional importance: the more conserved the amino acid sequence of a protein (or a segment of a protein) among species, the more critical the function of that protein (or protein segment). Highly conserved proteins (or protein segments) presumably carry out important functions that are disrupted by practically any amino acid substitution, and hence such substitutions are strongly selected against. Moreover, as also shown in Figure 18.17, several of the EDAR disease-causing mutations occur in the death domain, which further supports the functional importance of this domain. And, the 370A mutation also occurs in the death domain, which therefore suggests that the 370A mutation may be influencing EDAR function (i.e., activation of the NF-κB transcription factor) via the interaction between the EDAR death domain and the EDARADD protein.

How might one go about testing this hypothesis? A few years ago we were investigating the signal of selection on EDAR in East Asian populations, and two graduate students, Sean Myles and Jarek Bryk, came to me with a description of a cell line assay for EDAR function that had been used to show that some of the mutations in the EDAR death domain that were associated with ectodermal dysplasia did in fact abolish activation of NF-κB (Shimomura et al. 2004). Sean and Jarek were quite excited and wanted to try this assay with the 370A mutation, assuring me that it would be easy and quick, taking at most a month to carry out the experiment. I thought to myself, yeah, right, but gave them the go-ahead. Six months later—and with the help of many other people—they finally got reproducible (and publishable) results (Bryk et al. 2008); for those of you interested in the gory details of the assay, it is described in Box 18.1. But at least the results were worth waiting for, as Sean and Jarek were able to show that the 370A allele enhanced activation of NF-κB relative to the "normal" 370V allele (Figure 18.18). So now there was evidence to suggest a functional link between the 370A allele and the various associated phenotypic effects on hair, teeth, and so forth: namely, enhanced activation of the NF-κB transcription factor via the interaction between the EDAR death domain and the EDARADD protein.

Our satisfaction at having unraveled a piece of the puzzle was short-lived, however, as around the same time that we had gotten these results, a report appeared claiming that—using the same sort of assay that we had been using—the effect of the 370A allele was to *decrease,* rather than increase, activation of NF-κB (Fujimoto et al. 2008). Fortunately (for us), we soon became aware of a third study, also using a similar assay, that obtained the same results that we did (Mou et al. 2008). So, the consensus view is that the 370A allele does indeed enhance activation of NF-κB, and this view has been confirmed by subsequent studies. Incidentally, for those of you who are wondering how it can be that such opposite results can be obtained with the same assay, it is important to keep in mind that the assay utilizes living cells (see Box 18.1)—that is, it is not as simple as mixing together a few chemicals in a test tube and then seeing what happens. With living cells there are all sorts of confounding factors—what kind of cells are used, how the cells have been kept alive, what they are fed and how often, how fast they grow, and so forth, which make any assay based on living cells difficult to replicate even in the same lab (as witness the length of time it took us to get reproducible results), let alone across different labs.

BOX 18.1 ■ The Cell Line Assay to Determine the Effect of the EDAR 370A Mutation on NF-κB Activation



The logic behind the assay for EDAR activity associated with specific alleles. See text for details.

The assay is outlined in the accompanying figure and starts with an immortalized human cell line. The usual case with human cells is that if you place them in an appropriate culture medium containing all the goodies they need to survive and reproduce, they will happily grow and divide for awhile and then suddenly stop, grow old, and die, for no apparent reason—sort of like the bodies they came from! However, under certain special conditions cells can be coaxed into growing and dividing indefinitely, and such cell lines are said to be immortalized. Some cancer cells do this spontaneously, but a common way of accomplishing this with normal cells is to infect them with a particular virus, which can then induce the cells to become immortalized (as mentioned previously in Chapter 9).

Anyway, in addition to the immortalized cell line, the assay requires three different plasmids (circular pieces of DNA that can be introduced into the cells and contain the appropriate DNA segments so that they will be replicated along with the cell's own DNA as the cells grow and divide). The first plasmid contains a version of the *EDAR* gene with either the 370A or the 370V mutation, along with either the normal 375 allele, or with the 375H disease mutation. The 375H mutation was shown previously to eliminate NF-κB activation in this assay and hence serves as an important negative control; there should be no increase in NF-κB activation with any plasmid containing the 375H mutation, regardless of the allele at position 370. These plasmids are obtained by first introducing (cloning) the *EDAR* gene into the plasmid and then mutating the sites of interest to produce the desired mutations—if this sounds amazing, in a way it is, molecular biologists have come up with all sorts of cool ways to manipulate DNA (thus the term "genetic engineering"), and in fact you can buy off-the-shelf kits to do all of this! This plasmid has a strong promoter—that is, it has special DNA sequences that ensure that once the plasmid is in a cell, it will express high levels of EDAR RNA, which in turn will lead to high levels of EDAR protein inside the cell.

The second plasmid contains a gene that encodes an enzyme called luciferase, which converts a particular chemical (the enzyme's substrate) into another (the enzyme's product), with visible light also produced from this chemical reaction. As you might guess, the first luciferase gene was isolated from fireflies, although luciferases have been isolated from other creatures that luminesce, and these different luciferase genes use different substrates, which is a useful feature for the assay as described later. This plasmid also contains a special promoter that is activated only by NF-κB (remember, NF-κB is a transcription factor), and so luciferase will be produced only by the plasmid if NF-κB is present. Moreover, the amount of luciferase that is produced should be proportional to the amount of active NF-κB that is present—the more NF-κB, the more luciferase; the more luciferase, the more light produced when the luciferase substrate is added. The amount of light emitted can be measured with a device called a luminometer and thus reflects the amount of NF-κB—this is the key idea behind the assay.

The third plasmid consists of a luciferase gene from a different creature (in this case, from a type of coral called a sea pansy), which uses a different substrate than firefly luciferase and also produces a different wavelength (color) of light. The purpose of this plasmid is to serve as an important control in the assay, as described later.

So, now that we have all of our components, let's see how the assay works. First, the cell line is transfected with all three plasmids. To make sure that cells have indeed taken up the plasmids, the plasmids also express a protein that makes the cells resistant to a particular antibiotic. Treating the cells with the antibiotic after transfection thus kills off cells that have not taken up the plasmids. The antibiotic-resistant cells, which presumably have incorporated the plasmids, are then allowed to grow and divide. As they do so, they make lots of EDAR protein, which in turn leads to lots of NF-κB activation, which in turn leads to lots of firefly luciferase production. After enough time has gone by, the cells are lysed and then assayed for firefly luciferase activity by adding the appropriate chemical substrate and measuring the amount of light produced. The firefly luciferase reaction is then quenched by adding a chemical that stops the reaction, and the sea pansy luciferase activity is measured by adding the appropriate chemical substrate and again measuring the amount of light produced. Why do we need the sea pansy luciferase? Suppose we test cells that contain the EDAR 370A mutation and find more light (reflecting more NF-κB activation) than we find with cells that contain the EDAR 370V allele. Is this because the 370A allele itself activates more NF-κB than the 370V allele does? Or,

BOX 18.1 ■ (*Continued*)

is it because by chance there happened to be more cells in the assay when we tested the 370A allele than when we tested the 370V allele? Or, is it because the cells by chance happened to take up more plasmids when we added plasmids with the 370A allele than when we added plasmids with the 370V allele? Having the plasmid with sea pansy luciferase allows us to distinguish these sorts of experimental artifacts from a real difference between the 370A and 370V alleles on NF-κB activation. The amount of light produced by the sea pansy luciferase provides a baseline measurement that can be compared across different experiments: dividing the amount of light produced by the firefly luciferase assay by the amount produced by the sea pansy luciferase assay normalizes the results for any variation in plasmid uptake, number of cells, and so forth.

Since the cells used to carry out the assays still have their own *EDAR* gene that will activate NF-κB, you might wonder about the effect of this "endogenous" EDAR on the results. There is another important control that measures the influence of this endogenous EDAR on the results, and that is to transfect cells with the plasmid used to clone the *EDAR* gene, but without any inserted DNA (this is called the "empty vector" control), along with the other two plasmids in the assay. These cells are then grown and run through the assay, and the resulting light produced by the firefly luciferase thus reflects the background NF-κB activity, without any added effect of EDAR-containing plasmids. This background level of NF-κB activity is subtracted from the levels obtained in the experiments with EDAR-containing plasmids, in order to obtain the corrected level of NF-κB activity associated with the EDAR-containing plasmid. So, with these various controls, one can be reasonably confident that any variation in the level of NF-κB activity observed between cells transfected with the 370A allele versus cells transfected with the 370V allele can be attributed to the EDAR alleles and not to any experimental artifact (with the 375H disease mutation providing a further control, as this should result in essentially no NF-κB activity). Simple, no? No wonder it took 6 months to get reproducible results!

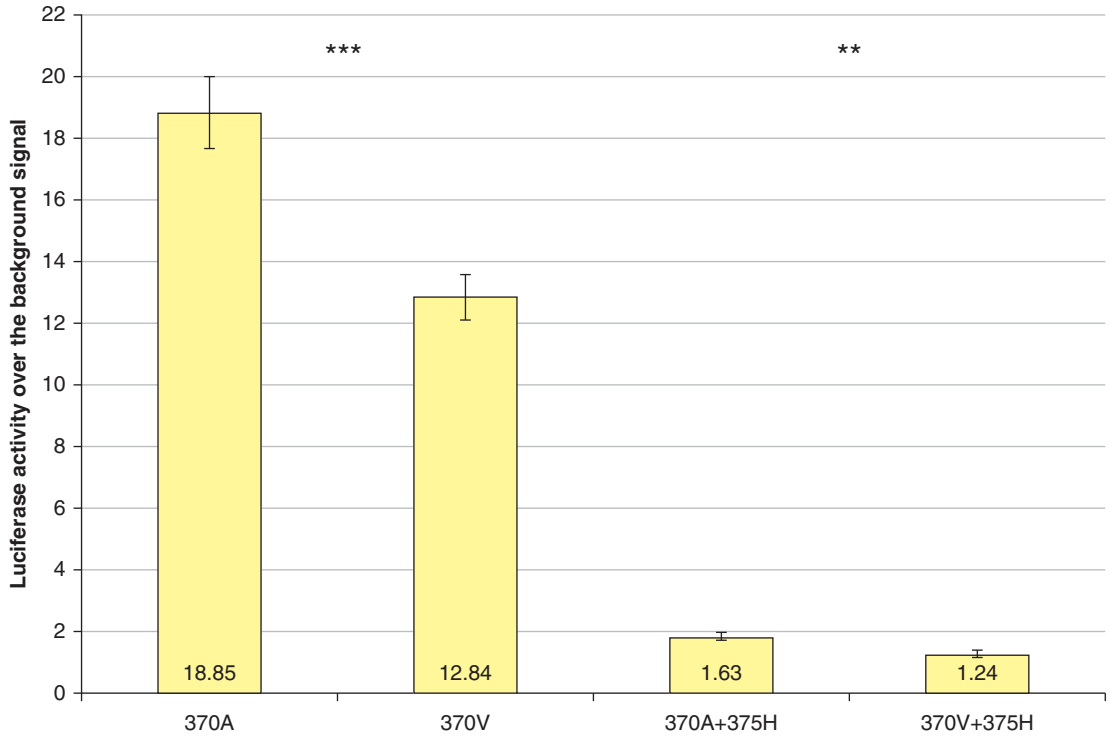


FIGURE 18.18

NF-κB activation in a cell line assay, as measured by luciferase activity (see Box 18.1 for details of the assay). The 375H mutation is a known disease-associated EDAR mutation that eliminates EDAR function and hence serves as a control to ensure that the assay works correctly. Note that the 375H mutation results in very low levels of activity, as expected, and that the 370A allele results in significantly higher levels of NF-κB activity than the 370V allele. The asterisks indicate significant differences between the adjacent columns as follows: $^{***}p < 0.001$; $^{**}p < 0.01$. Reprinted with permission from Bryk, J., et al., "Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation," *PLoS ONE* 3:e2209, 2008.

In addition to the functional experiments, we also estimated the time since fixation of the 370A mutation in the Han Chinese (Bryk et al. 2008). The approach is basically the same as that used to date the selection on FOXP2 and LP—namely, with strong selection, the 370A mutation will quickly rise in frequency and become fixed in the population along with the surrounding haplotype (as in Figure 18.1); recombination and new mutations will then break down the associated haplotype at a regular rate over time. So with reasonable estimates (or educated guesses) about the mutation and recombination rates, and the size of the population, you can get an estimate as to when selection drove the mutation to high frequency. And even though the 370A allele is not really fixed in Han Chinese (the 370A allele frequency is about 96% in Han Chinese), it is close enough that the approach works reasonably well. The result we obtained was that the 370A allele became "fixed" in the Han Chinese by 11,000 years ago, which makes sense in light of the finding that the 370A allele is also at high frequency in native Americans and, therefore, was probably already at high frequency in East Asians at the time of colonization of the New World around 15,000 years ago (as discussed in Chapter 16).

So now we have a mutation with a strong signal of selection that results in enhanced activation of a transcription factor, which (presumably) then has a number of phenotypic effects, including thicker hair and altered tooth morphology. But what is still missing is the answer to the "why" question: *why* was there such strong selection on the 370A mutation? There are three potential explanations to consider. First, one (or more) of the known phenotypic effects could indeed have a selective advantage in East Asia. But keep in mind that a selective advantage means that people with the phenotype conferred by the 370A allele would be having more surviving offspring than people with the 370V allele—and it is really hard to think of a reason why people with thicker hair and/or shovel-shaped incisors would enjoy such a selective advantage. The second possibility is that there is some additional—but as yet unknown—phenotypic effect of the 370A allele that conferred a selective advantage in East Asians. One admittedly speculative possibility involves disease resistance; NF-κB is known to activate some genes involved in the immune response to disease, so maybe there was some disease in East Asia for which the 370A allele increased resistance via enhanced NF-κB activation. The thicker hair and shovel-shaped incisors would then be an example of "phenotypic hitch-hiking"—they are secondary effects of the 370A allele that increased in frequency simply as a by-product of the selection on the really important (but as yet unknown) primary effect of the 370A allele. The third possibility would be sexual selection: people with thicker hair and/or shovel-shaped incisors were seen as more attractive, and hence preferential mating resulted in the increased frequency of the 370A allele. While sexual selection is a potentially interesting idea, it is hard to see how to test it.

While we (and others) were pondering these possibilities, further information came from additional studies of Denis Headon's mouse model with enhanced EDAR activity (Chang et al. 2009). Not only do these mice have thicker fur, it turns out they also have larger glands of various kinds—in particular, sebaceous glands (which secrete substances that lubricate and protect the body) and mammary gland networks—as well as enhanced rates of glandular secretions. This led to the proposal that these changes in gland size and secretion rate were the real target of selection on the 370A allele (assuming, of course, that the 370A allele in humans has the same effects on gland size and secretion rate as seen in the mice with enhanced mouse EDAR activity). The idea is that when modern humans got to East Asia about 25,000 to 30,000 years ago, the climate was much colder and drier than it is now. So, the 370A allele would have been selected for because the higher rates of glandular secretion would have helped protect exposed areas of the body. For example, Meibomian glands (which are sebaceous glands in the eyelids that secrete an oily substance that lubricates and protects the eyes) are enlarged in the enhanced EDAR mouse model, so if the same holds true for humans with the 370A allele, then it seems plausible that individuals with the 370A allele would have fared better in a cold, dry environment.

Moreover, there are other indications that East Asians may have adapted to a cold and dry environment. Ear wax, for example, comes in two varieties, wet and dry, and the dry form is found at high frequency in East Asia and is all but absent elsewhere. Dry versus wet earwax behaves as a Mendelian trait, and a few years ago the mutation responsible for dry earwax was found in a gene called *ABCC11* (Yoshiura et al. 2006). This gene encodes a protein that transports various substances into and out of cells, and the nonsynonymous mutation responsible for dry earwax seems to reduce transport efficiency. There is also a strong signal of recent positive selection associated with this mutation, dated to between 25,000 and 75,000 years ago (Ohashi et al. 2011). While this is a pretty broad range, it does overlap with the initial colonization of East Asia. And, since people with dry earwax also tend to sweat less, it has been suggested that the selection for the dry earwax mutation may have been to reduce sweating in a cold environment.

Another potential indication of adaptation to a dry environment is the epicanthic eye fold (i.e., the fold of skin over the lower part of the eye) that is characteristic of East Asian populations. At the moment

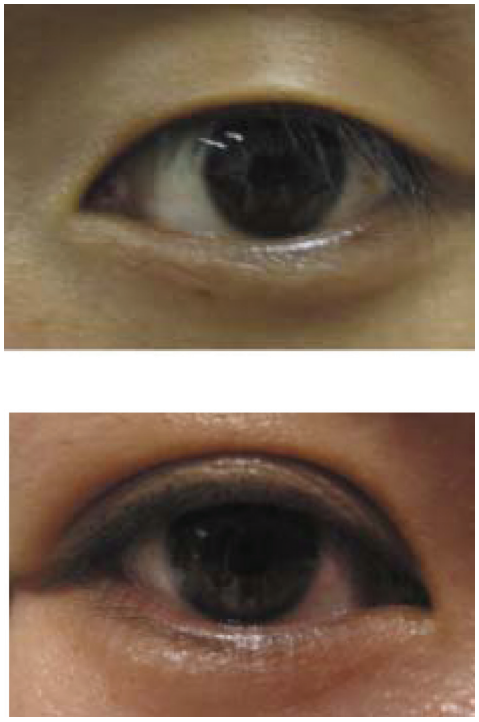


FIGURE 18.19

Single eyelid (top) compared to double eyelid (bottom). Reprinted with permission from Cho, M., and Glavas, I.P., "Anatomic properties of the upper eyelid in Asian Americans," *Dermatologic Surgery* 11:1736, 2009.

we don't know what mutation(s) cause the epicanthic eye fold to develop, nor do we know whether there was positive selection for this trait. But circumstantial support for the hypothesis that the epicanthic eye fold was an adaptation to an arid environment comes from the observation that similar eye folds tend to occur mostly in populations that live in arid/desert environments, such as some southern African groups that live in and around the Kalahari Desert. Another phenotypic trait that is mostly restricted to East Asian populations is the so-called "single eyelid" phenotype, in which there is an extra fat pad in the eyelid that results in a smooth eyelid without any creases, as compared to the "double eyelid" common in other populations, which lacks the fat pad and has a crease (Figure 18.19). This extra fat pad may also represent an adaptation to a cold environment, although do keep in mind that for both the epicanthic eye fold and for single eyelids, we don't know the genetic basis for these traits, nor do we even know whether there was indeed positive selection for these traits, so all of this is purely speculative. In fact, surgery to change single to double eyelids (called blepharoplasty, in keeping with the scientific tradition of using complicated terms whenever possible) is the most commonly performed plastic surgery for purely aesthetic purposes in Asia (and among Asian–Americans), and there is a wealth of discussion on the Internet as to which is more attractive, single or double eyelids—so this would seem to be a prime candidate for sexual selection!

Anyway, a plausible story can be told in which the primary selective force on the EDAR 370A mutation was for enhanced glandular secretions that helped protect skin and eye surfaces in the cold and arid environment that characterized East Asia some 20,000–30,000 years ago, and that the other traits associated with the 370A allele (i.e., thicker hair and shovel-shaped incisors) are likely to be secondary effects that hitch-hiked along with the primary effect. So far, so good—but a recent study has called this story into question. To gain further insights into what the 370A mutation does, the geneticist and rock star Pardis Sabeti and colleagues decided to make a humanized mouse that carries this mutation (analogous to the humanized FOXP2 mouse discussed in the previous chapter) and see what phenotypic effect(s) it has (Kamberov et al. 2013). And while the humanized 370A mouse shows some similarities to the previously discussed mouse model with enhanced (mouse) EDAR activity (such as thicker hair and more highly branched mammary gland networks), it also shows some intriguing differences. In particular, the humanized 370A mouse does not exhibit larger sebaceous glands (such as Meibomian glands) but does have more eccrine glands (the primary sweat glands), which would presumably enable more efficient cooling and thermoregulation via sweating. Sabeti and colleagues followed up on this finding by carrying out an association study in humans and found that individuals with the 370A allele do have more eccrine sweat glands than individuals with the "normal" 370V allele. They suggested that the selection on the 370A mutation may have been for more efficient sweating in response to a warm, humid environment, which characterized East Asia from about 40,000 to 30,000 years ago.

So what was the primary selective condition in East Asia on the 370A allele: the warm, humid environment from 40,000 to 30,000 years ago, or the colder, more arid environment from 30,000 to 15,000 years ago? Or, was it something else entirely? At the moment the jury is still out, and it is a bit frustrating that the situation is becoming murkier with additional studies, instead of clearer. Still, becoming more certain that we really don't know something, as opposed to thinking that we do know something that isn't actually true, is a form of progress (as Mark Twain said, "It ain't what you don't know that gets you into trouble, it's what you know for sure that just ain't so."). There is still lots more to be done, both with mouse models and with studies in humans, and even if we never come to a full understanding of the reason why there was selection

on the 370A mutation, for sure we'll learn more by trying to answer this question.

ANCIENT DNA

Just as with population history, ancient DNA is rapidly becoming an important source of new insights into recent selection in human populations. One source comes from archaic genomes; as mentioned in the previous chapter, maps of archaic ancestry in modern humans are being developed and improved (Sankararaman et al. 2014; Vernot and Akey 2014; Vernot et al. 2016), and these maps reveal not only deserts of archaic ancestry but also "islands," that is, regions of the genome with significantly more than the expected amounts of archaic ancestry. These islands indicate that some genes that we received from Neandertals and/or Denisovans (and perhaps other, as yet unknown archaic hominins) were subsequently selected for in human populations. This idea, known as "adaptive introgression," is intuitively appealing. After all, the ancestors of Neandertals and Denisovans left Africa several hundred thousand years ago, spread across Eurasia, and subsequently had to adapt to all of the new environments, climates, food sources, diseases, parasites, and so forth. And then, somewhere between 50,000 and 100,000 years ago or so, our ancestors left Africa, spread across Eurasia, and had to do the same thing all over again. So, if by interbreeding with archaic humans the early modern humans picked up genetic variants that were beneficial in Eurasia, this would have given the modern humans a head start toward adapting to the new conditions they encountered. Who knows, maybe this adaptive introgression was actually crucial to the survival of modern humans outside of Africa—we have no way of knowing this for sure, of course, but it's fun to think about.

Most of the genes that show the strongest signals of adaptive introgression are involved in the immune response to infectious disease (e.g., Abi-Rached et al. 2011), which is precisely what you would expect, given the devastating impact such diseases can have on human populations. But some genes with signals of adaptive introgression are more difficult to interpret. For example, populations that live at high altitude are genetically adapted to the low oxygen conditions (known as hypoxia) and it's been known for some time that a gene called *EPAS1*, which encodes a transcription factor (i.e., a protein that regulates the expression of other genes), contributes to the high altitude adaptation found in Tibetans (Simonson et al. 2010; Yi et al. 2010). What is rather astonishing, though, is the recent finding that the EPAS1 haplotype associated with high altitude adaptation in Tibetans seems to have come from Denisovans (Huerta-Sánchez et al. 2014)! Your guess is as good as mine as to what this means—maybe Denisovans were actually Yeti (abominable snowmen)? Anyway, investigating archaic genomes for important genetic traits that may have been contributed to us via adaptive introgression is a very active area of ongoing research, and undoubtedly there is a lot more to be learned.

Another way in which ancient DNA studies could potentially inform about recent selection is by providing unambiguous evidence as to the timing and dynamics of the spread of particular advantageous alleles. Although we discussed ways of dating the age of mutations back in Chapter 12, these dating methods inevitably come with large standard errors and, moreover, rely on assumptions about mutation rates and so forth that may or may not be reasonable. If one had a series of skeletal remains from the right place(s) and time(s), one could in theory directly follow the origin and spread of a particular adaptive mutation. However, such an approach requires a pretty decent sample size, and the issues of preservation and contamination discussed in Chapter 15 seemed to preclude ever getting enough authentic ancient DNA results to employ this approach. But the ongoing advances in ancient DNA methods—in particular, the recent realization that the petrous bone (part of the temporal bone of the skull, and one of the densest bones in humans) has amazingly good DNA preservation (Pinhasi et al. 2015)—has allowed analysis of hundreds of skeletal remains. For example, population genomicist David Reich and colleagues recently analyzed an astounding 230 skeletal remains from Eurasia, ranging in age from 8500 to 2300 years (Mathieson et al. 2015). Using capture enrichment to target specific SNPs of interest, they obtained genome-wide data for more than 1 million SNPs, which they then analyzed for signals of recent positive selection. The strongest candidates were the usual culprits, including lactase persistence (which they could show only increased strongly in frequency beginning about 4000 years ago), other genes associated with diet, and genes associated with immune response and with skin pigmentation. Of perhaps more interest is that they also investigated some complex traits (i.e., traits influenced by multiple genes and the environment) and found a significant signal of selection on height, which is all the more impressive given the difficulties in predicting height from associated SNPs (discussed in the last chapter). This pioneering study has certainly set the stage for further studies of selection using ancient DNA.

CONCLUDING REMARKS

Identifying, investigating, and (ultimately) understanding the signals of both species-wide and local

selection in the human genome remains a difficult enterprise. Demographic processes (population expansions/contractions) can mimic some signals of natural selection, and ascertainment bias can also have a big impact, as we saw in the case of the presumed balancing selection on the prion protein gene, discussed in the previous chapter. Nevertheless, while keeping these issues in mind, it is relatively straightforward to use genome scan approaches (in ancient as well as modern samples) to come up with lists of candidate genes; the problems arise with trying to then figure out which are the true candidates and which are false positives, and what to do next. Functional studies are always a good place to start and can sometimes lead to insights into the associated phenotypes. A major stumbling block, however, remains in identifying and testing hypotheses as to *why* selection has influenced the variation at a specific gene or genomic region. The ultimate goal is to understand the stories behind the adaptations that made us human and allowed us to successfully colonize more of the planet than any other creature, but at the same time you have to separate fact from fiction; storytelling based on pure speculation is not the same as finding out about the stories. Still, there have been some successes where we have learned a lot, even if we don't have the complete story, such as the examples of FOXP2, lactase persistence, and EDAR. Investigating the impact of selection on human evolution remains a hot topic, and the good news for students is that there is much more to be done, so there is lot of opportunity for the development of clever and creative new approaches.

CHAPTER 19

GENES AND CULTURE

One of the defining characteristics of humans is the degree to which we rely on culture in order to survive on this planet. By culture, I do not mean the original sense of the term, which had to do with cultivation of the soul or mind (with obvious parallels to agriculture) but rather the anthropological use of the word: namely, activities and behaviors (and associated goods and materials) that are learned or transmitted between individuals via observation, rather than innate or transmitted via genetic inheritance. And saying that culture is a defining characteristic of humans is not to deny that other creatures also have culture. Chimpanzees, for example, "manufacture" twigs to fish termites out of nests, or use stones to crack nuts, and, moreover, young chimps learn these behaviors by observing their elders, so these activities satisfy all the requirements of the anthropological definition of culture. But no other creature is as reliant on culture as we are—take away a chimpanzee's twigs or stones, so they can no longer fish for termites or crack nuts, and they'll still do just fine in the wild. But put one of us "modern" humans out in the wild without clothing, shelter, cars, cell phones, iPads, fast food restaurants, and so forth, and see how long we would last.

In this chapter, we will consider some aspects of the interaction between genes, evolution, and culture. One topic that will not be covered is how culture itself can evolve and change over time, as while the application of evolutionary principles to the study of culture is a very interesting topic, it is too broad to go into here. Instead, we will begin by considering the impact of culture on human evolution: in particular, is culture a barrier to human evolution, as has been often claimed? We will then see how culture can impact human genetic variation, both directly and indirectly. Finally, we will see some examples in which genetic analyses can be used to learn more about certain cultural practices—including a genetic approach to dating the origin of clothing (I kid you not!).

ARE HUMANS STILL EVOLVING?

It is not difficult to find writers stating quite bluntly that the effect of culture on human evolution is that of a barrier—that is, culture acts to prevent human biological evolution. For example, in an article by science writer Michael Balter on this question (Balter 2005), the geneticist and science popularizer Steve Jones was quoted as saying: "The central issue is what one means by 'evolving.' Most people when they think of evolution mean natural selection, a change to a different or better adapted state. In that sense, in the developed world, human evolution has stopped." And the anthropologist Ian Tattersall put it more succinctly: "Biologically, human beings are going nowhere." For sure, there is a lot of truth to this view. After all, biological evolution happens generally in response to some change in the environment that necessitates new genetic adaptations. Cultural change takes the place of new genetic adaptations because cultural changes can occur far more quickly and spread far more rapidly than genetic adaptations. For example, if the ozone layer continues to disappear and levels of ultraviolet radiation reach life-threatening levels, we'll most likely respond not biologically (by evolving thicker skin or the like) but rather culturally (by developing protective creams or clothing, or placing shields over cities, or moving cities underground). So, it is easy to see that culture can indeed act as a barrier to biological evolution.

However, at the same time, cultural traits can induce profound changes that—directly or indirectly—influence human genetic variation and, ultimately, biological evolution. In the next sections, we will see some examples of such cultural traits and their impact on genetic variation. In fact, in contrast to the popular view that culture is a barrier to human evolution, there is an opposing view that holds that culture has actually accelerated human evolution. That is, some of

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

the most important influences on recent human evolution involve cultural changes, such as the development of agriculture and associated increases in population density, the formation of cities and nation-states, the industrial and technological revolutions, and so forth, all fueled by the rapid demographic growth of human populations. Among the lines of evidence cited in support of this view is the finding that there are more signals of recent selection in our genome than of older selection events (Hawks et al. 2007), based on the sorts of studies discussed in the previous chapter. However, while I am very sympathetic to the view that culture has indeed influenced human evolution, and perhaps even accelerated the rate at which humans are evolving, I do not think that scans for signals of selection in our genome can be used to argue this point, because there is an inherent bias in such studies. Recent selection is much easier to detect than ancient selection in genome scans, because the principal signal of selection in such scans (LD and/or extended haplotype homozygosity) breaks down over time and will disappear entirely after 20,000 years or so. There could easily have been just as much, if not more, selection operating on humans in the distant past, but with current methods it is much more difficult to detect the signal of older selection events in our genomes. So, maybe there has been more selection on humans recently as a consequence of culture, or maybe not—with current data we can't tell. But what we certainly can do is identify particular examples of genes whose variation has been directly impacted by a particular cultural trait, as well as aspects of human genetic variation that have been indirectly influenced by cultural practices, and that is what we will turn to now.

Genetic Variation Can Be Directly Influenced by Cultural Practices

This section will be fairly short, because we've already discussed (in the previous chapter) the "poster child" for a genetic trait whose variation has been directly influenced by cultural practices, namely, lactase persistence. To briefly recapitulate, in populations that drink milk as a source of nutrition, there has been strong selection favoring genetic variants that allow humans to digest lactose (the major sugar present in mammalian milk) into adulthood. Thus, genetic variation for this trait has been directly influenced by the cultural practice of drinking milk. Another example of a cultural practice related to diet that has influenced variation for a particular gene involves salivary amylase, which is an enzyme found in our saliva that helps break down starch for digestion. There is variation in the number of genes for salivary amylase among different populations (Perry et al. 2007), and this variation is directly correlated with salivary amylase activity (i.e., the more genes you have, the more salivary amylase you make, and the quicker and easier it is for you to digest starch). Moreover, populations with diets that are traditionally high in starch (e.g., all agricultural groups) have significantly more salivary amylase gene copies than do populations with diets that are low in starch (e.g., rain forest hunter-gatherers or Siberian reindeer herders), suggesting that selection has acted to increase the number of salivary amylase genes in response to the starch-rich diet that accompanied the advent of agriculture. Interestingly, this trend is not limited to humans—it turns out that dogs also have more amylase genes than their wild progenitors, wolves, and this increase was selected for during dog domestication, presumably in response to the diet of dogs becoming more starch-rich because of their association with humans (Axelsson et al. 2013). So, in this respect, dogs can be thought of as carbo-loaded wolves. And, just like humans, dogs suffer from starch-rich diets by exhibiting higher rates of type 2 diabetes and other metabolic disorders—it truly is a dog's life!

Other examples of genes that exhibit variation that has been directly impacted by selection are not so easy to come by—not necessarily because they are rare but because (as we saw in Chapter 18) while it is relatively straightforward to identify candidate genes that show signatures of local adaptation, it is much harder to figure out why a particular gene shows a particular signal of selection. For most of these candidate genes we don't really have a clue as to what is behind the signal of selection, but it seems reasonable to suppose that as we get better at figuring this out, we will find more examples of cultural traits having a direct impact on the genetic variation at a particular gene.

Genetic Variation Can Be Indirectly Influenced by Cultural Practices

In addition to directly impacting the variation at particular genes, cultural practices can also have an indirect impact on human genetic variation. Perhaps the best examples involve social practices that are sex-biased. By comparing patterns of maternally inherited mtDNA and paternally inherited NRY variation, we can discern the influence of sex-biased practices on genetic variation. Indeed, the potential impact of sex-biased social practices on human genetic variation was inferred in one of the first comparative studies of mtDNA and NRY variation across human populations. This seminal study, published in 1998 and carried out by the legendary Luca Cavalli-Sforza and colleagues (Seielstad et al. 1998), apportioned mtDNA, NRY,

TABLE 19.1 ■ Apportionment of variation for autosomal DNA, mtDNA, and Y chromosomal DNA in human populations[a]

Genetic system	Within populations	Within continents	Between continents
Autosomes	85.6	5.7	8.8
mtDNA	81.4	6.1	12.5
Y chromosome	35.5	11.8	52.7

[a]Data from Seielstad, M., et al., "Genetic evidence for a higher female migration rate in humans," *Nature Genetics* 20:278, 1998.

and autosomal DNA variation into within-population, within-continent, and between-continent components (Table 19.1). The biggest of these components for mtDNA and autosomal DNA was the within-population component, clocking in at around 80–85% (corresponding to F_{ST} values of around 0.15–0.20, as discussed previously in Chapter 10 and shown in Table 10.2). However, for the NRY the within-population component was only 35% (corresponding to a whopping F_{ST} value of 0.65). Moreover, a comparison of genetic (F_{ST}) distance with geographic distance showed a highly significant correlation for all three genetic systems, but genetic distance increased much more rapidly with geographic distance for the NRY than for mtDNA or autosomal DNA (Figure 19.1). In other words, populations separated by the same geographic distance show much bigger genetic distances for the NRY than for mtDNA or autosomal DNA.

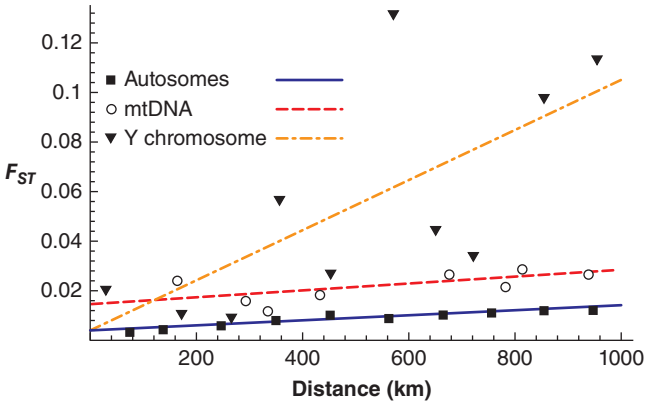


FIGURE 19.1

Genetic (F_{ST}) distances versus geographic distance for pairs of populations, for mtDNA, Y chromosome, and autosomal DNA. Note that the F_{ST} distance for the Y chromosome increases much faster with increasing geographic distance than do the F_{ST} distances for either mtDNA or the autosomes. Modified with permission from Seilestad, M.T., et al., "Genetic evidence for a higher female migration rate in humans," *Nature Genetics* 20:278, 1998.

How to explain this huge discrepancy between the NRY and the other genetic systems? Cavalli-Sforza and colleagues considered several possible explanations, including selection on the NRY, higher male mortality, or **polygyny** (mating systems in which some men have more than one wife, with the result that some men end up without any wives), all of which could in theory reduce the effective size for the NRY and hence increase genetic differences between populations. However, these factors are insufficient to account for the large differences between the NRY and the other genetic systems shown in Table 19.1 and Figure 19.1. Instead, the results are best explained in terms of migration: namely, the genetic results are compatible with higher rates of female than male migration. This may not seem like a very plausible explanation, because as I wrote back in 1998 in a commentary accompanying the Cavalli-Sforza study (Stoneking 1998), when we think of migration, the

> … image that often comes to mind is that of the intrepid explorer leading the way into the unknown, or the conquering hero subjugating the denizens of distant lands (think Marco Polo, Alexander the Great, Genghis Khan or Attila the Hun and you get the idea).

But while long-distance migration may indeed be male-dominated, it pales in importance when compared to the female migration that occurs when men and women get married. It turns out that the vast majority of human societies are **patrilocal**, meaning that when a man marries a woman from a different location, the woman moves to the residence of the man. With patrilocality, every generation females are moving around while males stay put. The genetic consequences thus are that mtDNAs are moving around much more between populations than are Y chromosomes, thereby leading to smaller genetic differences between populations for mtDNA. Conversely, lower rates of male migration will lead to bigger genetic differences between populations for the NRY. Cavalli-Sforza and colleagues estimated that an average rate of female migration that is eightfold higher than male migration would be sufficient to account for the observations in Table 19.1 and Figure 19.1.

There is an obvious test of the hypothesis that bigger differences among populations for the NRY than for the mtDNA (or autosomal DNA) reflect patrilocality, and that is to examine patterns of mtDNA/NRY variation in **matrilocal** groups. These are groups in which the male moves to the residence of the female after marriage. If relative rates of female versus male

migration are indeed having an impact on patterns of genetic variation, then in matrilocal groups we should expect to see the opposite pattern, namely, bigger differences among populations for the mtDNA than for the NRY. A matrilocal residence pattern is relatively rare among human societies but there are some. I was fortunate that at around the time that Cavalli-Sforza's study came out, I was contacted by a Japanese researcher, Hiroki Oota, who had participated in an anthropological survey in 1996 of some elusive groups in northern Thailand called hill tribes. Some hill tribes are patrilocal while others are matrilocal, and the survey had collected samples for DNA analysis; Hiroki was keen to bring the DNA samples to my laboratory in order to analyze mtDNA and NRY variation in them. It turned out that the hill tribes were ideal for this sort of study, as there were three patrilocal and three matrilocal groups sampled, all from the same general geographic region (northern Thailand), all practicing the same subsistence strategy (slash and burn agriculture), and all speaking closely related Sino-Tibetan languages. Thus, several factors that can potentially influence patterns of genetic variation (i.e., geography, subsistence, and language) are the same for all of the groups, making it more likely that any differences in patterns of mtDNA/NRY variation between the matrilocal and patrilocal hill tribes do reflect the different residence patterns rather than something else. In other words, it's not like we had to compare matrilocal groups from the Amazon to patrilocal groups from Siberia or something like that, where there would be all sorts of potential confounding influences on mtDNA and NRY variation.

If patrilocal versus matrilocal residence pattern is indeed influencing mtDNA versus NRY variation in the hill tribes, then there are two predictions we can make and test. Note that we can't directly compare mtDNA variation to NRY variation, because different methods were used to assay mtDNA variation (by sequencing the first hypervariable segment of the control region) versus NRY variation (by genotyping several STR loci), and these different molecular methods will influence diversity estimates. Moreover, the different mutation rates for mtDNA versus the Y chromosome also complicate any attempt to directly compare the two. But we can compare mtDNA variation in matrilocal versus patrilocal groups, and similarly we can compare NRY variation in matrilocal versus patrilocal groups, without running into this problem of different molecular methods or different mutation rates. The first prediction that we can test is that mtDNA diversity should be lower in matrilocal groups than in patrilocal groups, while NRY diversity should be lower in patrilocal groups than in matrilocal groups (Figure 19.2). This is because in matrilocal groups, females (and hence mtDNA genomes) are staying in the group they were born in, whereas in patrilocal groups they are moving around between groups. This has the effect of lowering the effective population size for mtDNA in matrilocal groups relative to patrilocal groups, and hence genetic diversity for mtDNA is also expected to be lower in matrilocal groups than in patrilocal groups. And, based on similar logic, NRY diversity should be lower in patrilocal groups than in matrilocal groups. And the results? Rather to our astonishment (because things rarely work out the way you expect them to), this prediction was exactly fulfilled (Oota et al. 2001): mtDNA diversity was significantly lower in the matrilocal groups than in the patrilocal groups, whereas NRY diversity was significantly lower in the patrilocal groups than in the matrilocal groups (Figure 19.3).

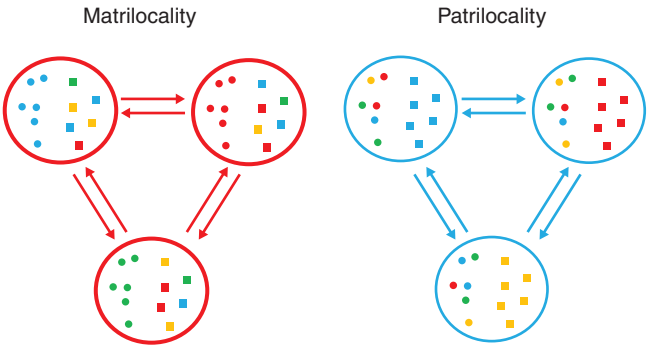


FIGURE 19.2

Predicted impact of residence pattern (matrilocality vs. patrilocality) on mtDNA/NRY diversity and divergence. Small circles denote mtDNA types and small squares denote NRY types; larger red circles denote matrilocal groups and larger blue circles denote patrilocal groups, and arrows indicate gene flow.

The second prediction, if residence pattern is influencing mtDNA and NRY variation, is that the mtDNA divergence between groups should be bigger for matrilocal groups than for patrilocal groups, whereas NRY divergence between groups should be bigger for patrilocal groups than for matrilocal groups. If this is not immediately obvious, see Figure 19.2: lower effective population sizes enhance the effects of genetic drift, thereby resulting in bigger genetic differences between groups (as also explained back in Chapter 5, in the section on genetic drift), and so this is expected to be the case for the mtDNA in matrilocal groups and for the NRY in patrilocal groups. And the results? Again, this prediction was exactly fulfilled (Figure 19.4). The inescapable conclusion: residence pattern is indeed having a profound, albeit indirect, influence on mtDNA and NRY variation in human populations. Moreover, these results further support the claim by Cavalli-Sforza and colleagues

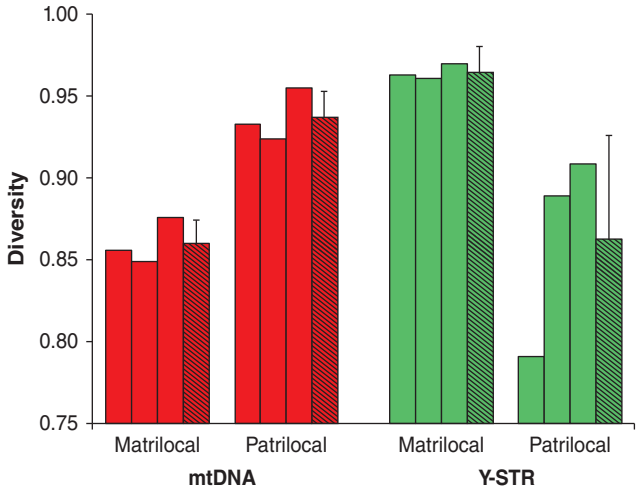


FIGURE 19.3

mtDNA and Y chromosome (Y-STR) diversity within matrilocal and patrilocal hill tribe groups. Each bar shows the diversity for an individual group; hatched bars show the averages for the three matrilocal or three patrilocal groups, respectively. Reprinted with permission from Oota, H., et al., "Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence," *Nature Genetics* 29:20, 2001.

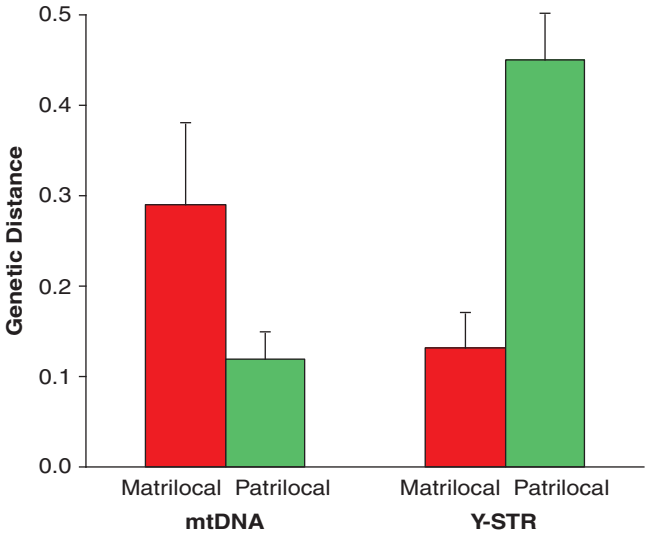


FIGURE 19.4

Genetic distances based on mtDNA and Y chromosome data for the matrilocal and patrilocal hill tribe groups. Reprinted with permission from Oota, H., et al., "Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence," *Nature Genetics* 29:20, 2001.

(Seielstad et al. 1998) that the bigger genetic differences between human populations for the NRY than for mtDNA reflect increased female migration as a consequence of patrilocality.

In some respects, we were fortunate to get such clear-cut results, as that doesn't happen very often in this business. Indeed, subsequent studies of other matrilocal groups have sometimes found similar results (e.g., Bolnick et al. 2006) but other times have not (e.g., Kumar et al. 2006)—probably because residence pattern is just one of many factors that can influence patterns of genetic variation, and how influential residence pattern is compared to these other factors will vary from situation to situation. Moreover, residence pattern is not an absolute, fixed trait but can vary over time. Still, most studies of patrilocal groups do find an association between residence pattern and mtDNA/NRY variation—even chimpanzees, which have strongly female-biased dispersal (i.e., are extremely patrilocal, probably even more so than most human groups), show the predicted patterns of mtDNA/NRY variation (Langergraber et al. 2007). In addition, even though a group may have clearly prescribed social practices, sometimes other circumstances will lead to exceptions to the rule. That was one of the most important lessons I've learned from fieldwork—for example, among groups in the Western Province of Zambia, the customary practice is for an individual to belong to the mother's clan (so, these groups are **matrilineal**, but not matrilocal). However, when gathering information about the individuals we sampled, sometimes they would claim membership in their father's clan (usually because the father's clan was dying out and so needed more members), and sometimes they would claim membership in an entirely different clan (e.g., because their family had moved to the traditional territory of that clan, and that's where they had grown up). The take-home message: human social rules are flexible and can be adapted to meet different circumstances as needed—but if you spend all your time in the laboratory or at the computer and don't get out into the field and talk to the people, then you may not appreciate the extent to which this can happen.

There have been additional studies of global patterns of mtDNA and NRY variation, and not all of them have found the large differences between F_{ST} values for mtDNA versus the NRY that Cavalli-Sforza and colleagues observed in their seminal 1998 study (e.g., Wilder et al. 2004). These studies suggest that patrilocality has not had a significant influence on global patterns of mtDNA and NRY variation. However, most of these studies sampled a limited portion of the mtDNA and NRY variation in a small number of populations (which, to be fair, is also true of the study by Seielstad et al. 1998). Recently, a high-resolution study (Lippold et al. 2014) carried out in my laboratory of complete

mtDNA genome sequences and about 500 kb of NRY sequence from each of the 600 or so males in the Human Genome Diversity Panel, which come from 52 populations (see Chapter 9), found F_{ST} values of 0.25 for mtDNA and 0.36 for the NRY—not as extreme as those found in the seminal 1998 study but still supporting the idea that genetic differences are indeed bigger among global human populations for the Y chromosome than for mtDNA. However, there was also significant regional variation in patterns of Y chromosome versus mtDNA differentiation, so to speak of global patterns may not have much meaning. Nonetheless, many studies of mtDNA and NRY variation carried out at a much more local scale (and in patrilocal groups) have found bigger F_{ST} values for the NRY than for mtDNA. To repeat: the inescapable conclusion is that the social practice of whether the man or the woman moves to the residence of the other after marriage has had a profound—albeit indirect—impact on patterns of human genetic variation.

Residence pattern is not the only social practice that has influenced patterns of human genetic variation. Another example is the caste system in India, which (until recently) existed as a hierarchical system of socially defined groups that individuals were born into. Although there were numerous castes, which used to govern religious practices, access to education, occupational opportunities, and so forth, they were grouped into several different levels according to status, and these different levels also governed marriage. Marriages within the same status level were strongly encouraged; occasionally, males were permitted to marry females with a lower status (in which case, the children usually obtained the status of the father), but lower-ranking males were essentially never permitted to marry a higher-ranking female. Note that the caste system thus would be predicted to have different effects on mtDNA versus NRY variation: genetic differences between status levels should be bigger for the NRY than for the mtDNA, because of no male "migration" between status levels compared to some female "migration" (analogous to the effects of patrilocality). Moreover, the female mobility between status levels should lead to a correlation between mtDNA differentiation and status level, while the lack of any male mobility between status levels should lead to no association between NRY differentiation and status level (i.e., in the absence of male gene flow between status levels, genetic drift and random mutations will be the only influence on NRY variation). And indeed, the genetic evidence bears out this prediction (Bamshad et al. 1998): genetic distances between status levels are about 10 times higher for the NRY than for the mtDNA, and there is a general clustering of groups by status level in a neighbor-joining tree based on mtDNA variation but no such clustering in the tree based on NRY variation (Figure 19.5). The conclusion: the social practice of the caste system has (indirectly) influenced patterns of mtDNA versus NRY variation.

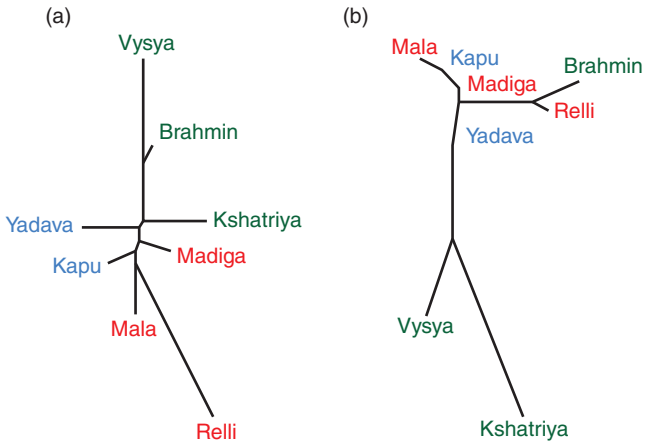


FIGURE 19.5

Genetic relationships of caste groups from India. Shown are neighbor-joining trees constructed from genetic distances based on (a) mtDNA HV1 sequences and (b) Y-STR loci. Green, upper castes; blue, middle castes; red, lower castes. Note that genetic distances are associated with caste hierarchy in the case of mtDNA but not for the Y-STR loci. Modified with permission from Bamshad, M.J., et al., "Female gene flow stratifies Hindu castes," *Nature* 395:651, 1998.

USING GENETIC ANALYSES TO LEARN MORE ABOUT CULTURAL PRACTICES: AGRICULTURAL EXPANSIONS

Having established that cultural practices can indeed influence genetic variation, we can turn this around and use genetic analyses to learn something more about particular cultural practices. We will proceed in this section as we did in the previous sections, namely, by going through some specific examples to illustrate this point. Let's start with one of the most important developments that humans ever came up with, namely, agriculture. As shown in Figure 19.6, plant (and animal) domestication happened at several different times at several different locations and then spread from these centers sometimes across quite large distances (e.g., from the Middle East across Europe, from China across Southeast Asia, from western Africa across nearly all of sub-Saharan Africa). A key question that then arises is: how did this spread occur? Did foraging groups see neighboring groups practicing agriculture, say to themselves, hey, look what those guys are doing, and then switch to agriculture? If so, then agriculture spread via **cultural diffusion**, without any substantial migration of people. Or, did

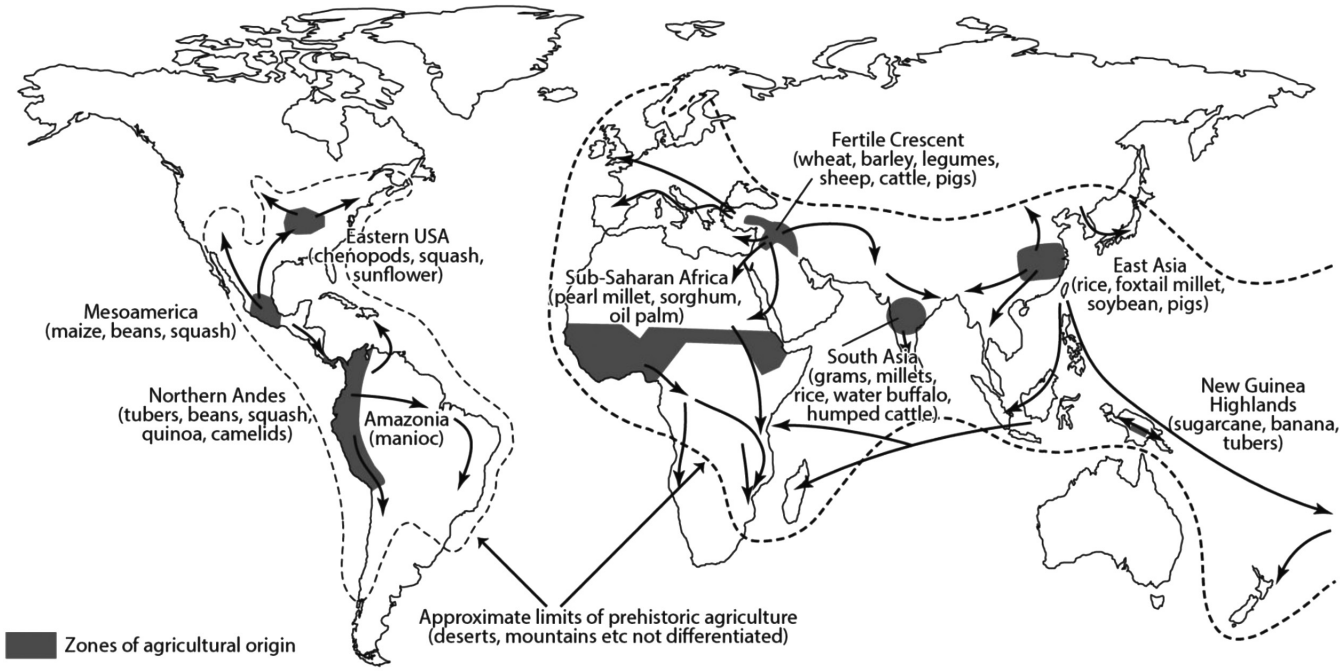


FIGURE 19.6

Estimated places and times associated with the domestication of different plants and animals. Arrows show hypothetical directions of the spread of agriculture, and the dashed lines indicate the approximate limits to the spread of agriculture in prehistoric times. Provided by, and reprinted with permission from, Peter Bellwood.

agricultural groups increase in size and expand geographically, taking over the areas most suitable for growing their crops and animals, and either assimilated any foraging groups they encountered along the way, replaced them without any admixture, or drove them into less hospitable territories? If so, then agriculture spread by **demic diffusion**, meaning that the spread of farming was accompanied by the spread of farmers. Genetics offers a way to distinguish between these two possible explanations: with cultural diffusion, agriculture spreads without any significant movement of farmers or their genes, while with demic diffusion, agriculture spreads primarily via farmers moving with their genes.

Let's consider by way of example what has probably been the most-studied agricultural expansion, namely, the spread of farming from the Fertile Crescent to Europe, also known as the Neolithic transition. Archaeological evidence documents the rise of agricultural societies around 11,000 years ago in western Asia, with farming spreading rapidly to southeastern Europe by 9000 years ago and then across all of mainland Europe by 5000 years ago, but of course the archaeological evidence doesn't tell you whether this rapid spread of farming across Europe was via cultural or demic diffusion. The first genetic evidence to address this question (Ammerman and Cavalli-Sforza 1984) involved principal components (PC) analysis of classical markers—those of you with good memories may recall that we briefly mentioned this way back in Chapter 10, when discussing the interpretation of synthetic maps based on PC analysis. To help refresh your memory, Figure 19.7 shows again the synthetic map for the first PC, which shows a cline (i.e., gradient in allele frequency changes) across Europe that is significantly correlated with the spread of agriculture across Europe. Luca Cavalli-Sforza and colleagues argued that this cline in the first PC was strong evidence in favor of the demic diffusion hypothesis. However, while this may certainly be the case, the evidence is purely circumstantial—there is nothing to directly link the synthetic map of the first PC to the spread of agriculture. In addition to the spread of agriculture, there have been several important demographic events in Europe that involved a similar spread from southeast to northwest. These include the initial colonization of Europe by modern humans some 45,000 years ago, as well as the retreat of populations to certain refugia (especially in southern Europe) during the Last Glacial Maximum, which lasted from about 27,000 to about 16,000 years ago, followed by population spread as the glaciers retreated. So, the cline in the first PC could in principle reflect either of these events, or some other entirely different population movement, or a combination of population movements, or even no population movement at all (e.g., isolation by distance, with the

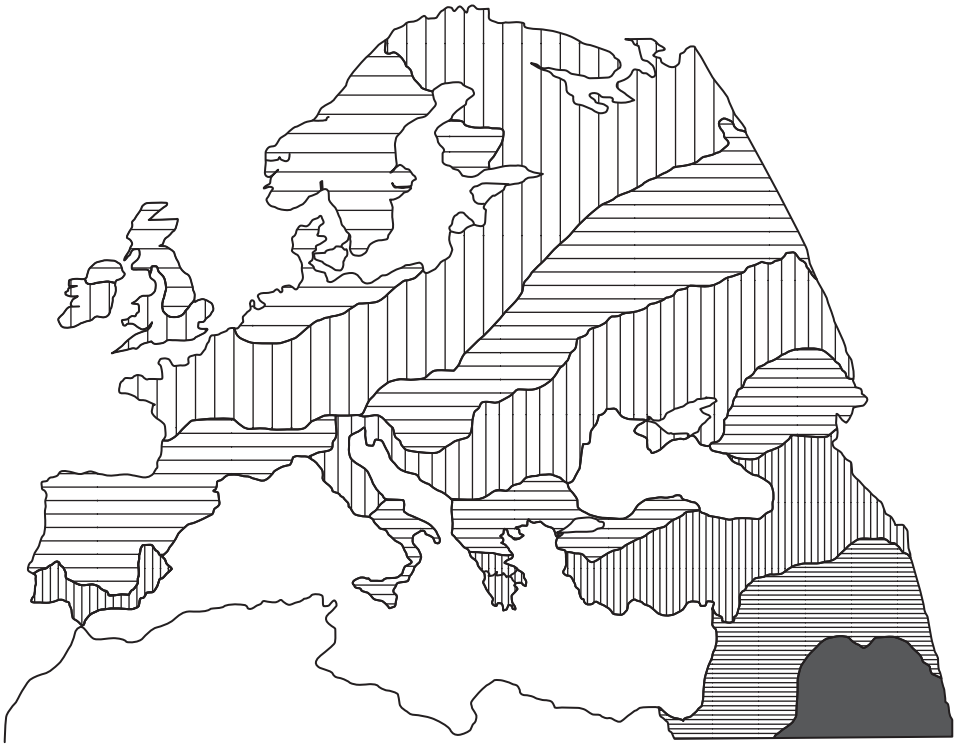


FIGURE 19.7

Synthetic map of PC1 values for Europe, based on classical genetic markers. Modified with permission from Cavalli-Sforza, L.L., "Genes, peoples, and languages," *Proceedings of the National Academy of Sciences USA* 94:7719, 1997.

amount of gene flow between populations inversely related to the geographic distance separating them), rather than the spread of agriculture.

Beginning in the 1990s, numerous studies used DNA evidence (mtDNA, NRY, and autosomal DNA loci, especially STR markers) to try to address this question of cultural versus demic diffusion. However, these studies generated more confusion than clarity, as estimates of the genetic contribution of Neolithic farmers to the present-day European population ranged from less than 15% to more than 70%. Early studies of mtDNA variation focused on analyses of the first hypervariable segment of the control region (HV1) and/or RFLP markers diagnostic for particular haplogroups. These studies tended to focus on phylogeographic analyses, that is, figuring out when and where particular haplogroups arose, and then taking a rather simplistic view that haplogroups older than about 10,000 years represent a pre-Neolithic contribution to the European populations (Figure 19.8). But, as was emphasized in Chapter 12, ages of haplogroups do not equate to ages of populations, and a haplogroup that arose more than 10,000 years ago could easily have been brought to Europe much more recently (e.g., by Neolithic farmers). There also did not seem to be any evidence of clinal variation in mtDNA, which some took as an argument against a strong influence of any migration of farmers across Europe—although NRY studies (Figure 19.9) as well as additional autosomal DNA studies (e.g., Chikhi et al. 1998) did continue to find evidence of clinal variation. A possible explanation for this apparent discrepancy between mtDNA and other markers is that there was a Neolithic expansion of farmers from western Asia that involved mostly males, who then interbred with local females, thereby accounting for the apparent lack of a major Neolithic contribution to European mtDNAs but evidence for a substantial Neolithic contribution to the rest of the European genome. However, this is purely speculative.

At any rate, the relative importance of Neolithic versus pre-Neolithic contributions to the contemporary European gene pool remained contentious for many years. Much of the debate centered around how different genetically any putative Neolithic farmers coming from western Asia were from the resident European hunter-gatherers; the more similar they were, the more difficult it becomes to distinguish their genetic contributions. The debate was complicated by the fact that there are no extant hunter-gatherer populations in Europe, so some researchers have used population isolates—such as the Basque or Sardinians—as proxies for the pre-Neolithic populations of Europe, with mixed results. While the Basque and Sardinians do differ (somewhat) genetically from other European populations, it is not clear whether this is because they actually do have a higher pre-Neolithic genetic

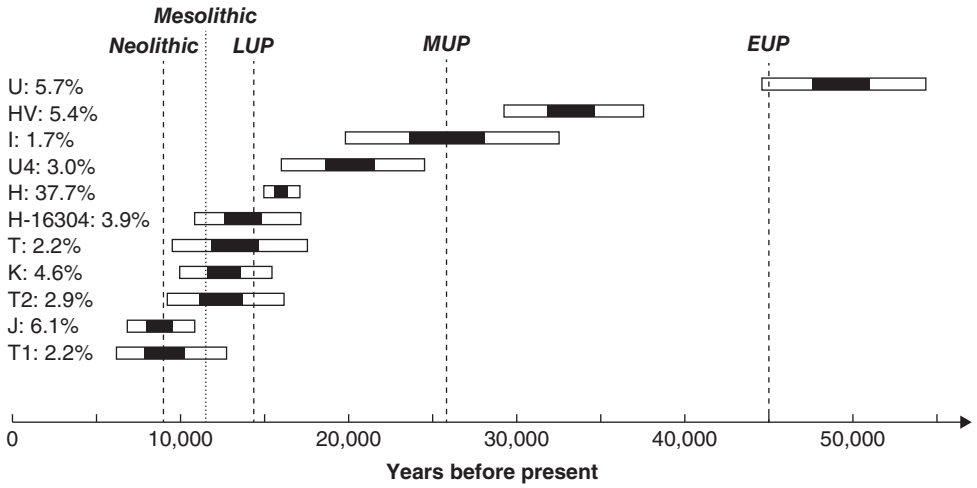


FIGURE 19.8

Ages of mtDNA haplogroups, showing that most haplogroups are older than the Neolithic—but, contrary to how some have interpreted this observation, this does not mean that they did not expand into Europe with Neolithic farmers. Modified with permission from Richards, M., et al., "Tracing European founder lineages in the Near Eastern mtDNA pool," *American Journal of Human Genetics* 67:1251, 2000.

component than other European populations, or whether they had just as much of a Neolithic genetic contribution as other European populations but subsequently experienced more genetic drift due to small population size and isolation.

This is a situation in which ancient DNA would seem to be the obvious route to go (assuming, of course, that one can overcome all the technical obstacles and contamination issues discussed in Chapter 15): direct comparisons of DNA from pre-Neolithic and early Neolithic skeletons should give us the answer. And indeed, new findings from ancient DNA are doing just that and are painting a more coherent picture. The first such studies focused on mtDNA, because of the previously discussed advantages of mtDNA for ancient DNA (in particular, the higher copy number), and a summary of the results is in Figure 19.10. The picture that is emerging is that all mtDNAs from pre-Neolithic hunter-gatherers belong to one of several subclades of haplogroup U, while skeletons from the earliest Neolithic farmers in Europe have mtDNAs from other haplogroups, with hardly any belonging to haplogroup U. mtDNA types from presumed hunter-gatherers that date from the early Neolithic (so, after the arrival of farmers) are still predominantly haplogroup U but do show an increased frequency of other mtDNA haplogroups relative to pre-Neolithic hunter-gatherers, suggesting genetic interactions with the early farmers. Similarly, skeletons from later Neolithic farmers show more haplogroup U mtDNA types than do those from the earliest Neolithic farmers, also suggesting genetic interactions with hunter-gatherers. Today, haplogroup U ranges in frequency from 5% to 25% across Europe, and this may be the best estimate yet for the contribution of pre-Neolithic groups to the contemporary European mtDNA gene pool. Moreover, the absence of what are today the most common European haplogroups (H, J, K, and T) in pre-Neolithic Europeans argues strongly for a Neolithic introduction of these haplogroups. It is also worth pointing out that haplogroup U occurs at frequencies of about 20% in Basque groups and at 5–25% in Sardinian groups, which does not suggest a particularly large "pre-Neolithic" genetic component in their maternal gene pools.

More recently, several genome-wide data sets from ancient Europeans have appeared. The population genomicist Mattias Jakobsson and colleagues used next-generation sequencing methods to obtain 27–97 million base-pairs of sequence from the skeletal remains of three hunter-gatherers and one farmer, all around 5000 years old (Skoglund et al. 2012). The remains all came from Sweden, accounting for their relatively good preservation (about 2–6% of the endogenous DNA was human). Because of the relatively small amount of sequence obtained ("only" a few million base-pairs!), there wasn't enough overlapping sequence to analyze all of the data together, so Jakobsson and colleagues relied on some neat statistical tricks to compare each sequence separately to data from contemporary populations in a PC analysis (thereby maximizing the number of SNPs in each comparison) and then combine the separate PC analyses into one plot. The results (Figure 19.11) are quite interesting: the ancient farmer is closest genetically to contemporary southern European populations, while the three ancient hunter-gatherers are genetically quite distinct from any contemporary European population

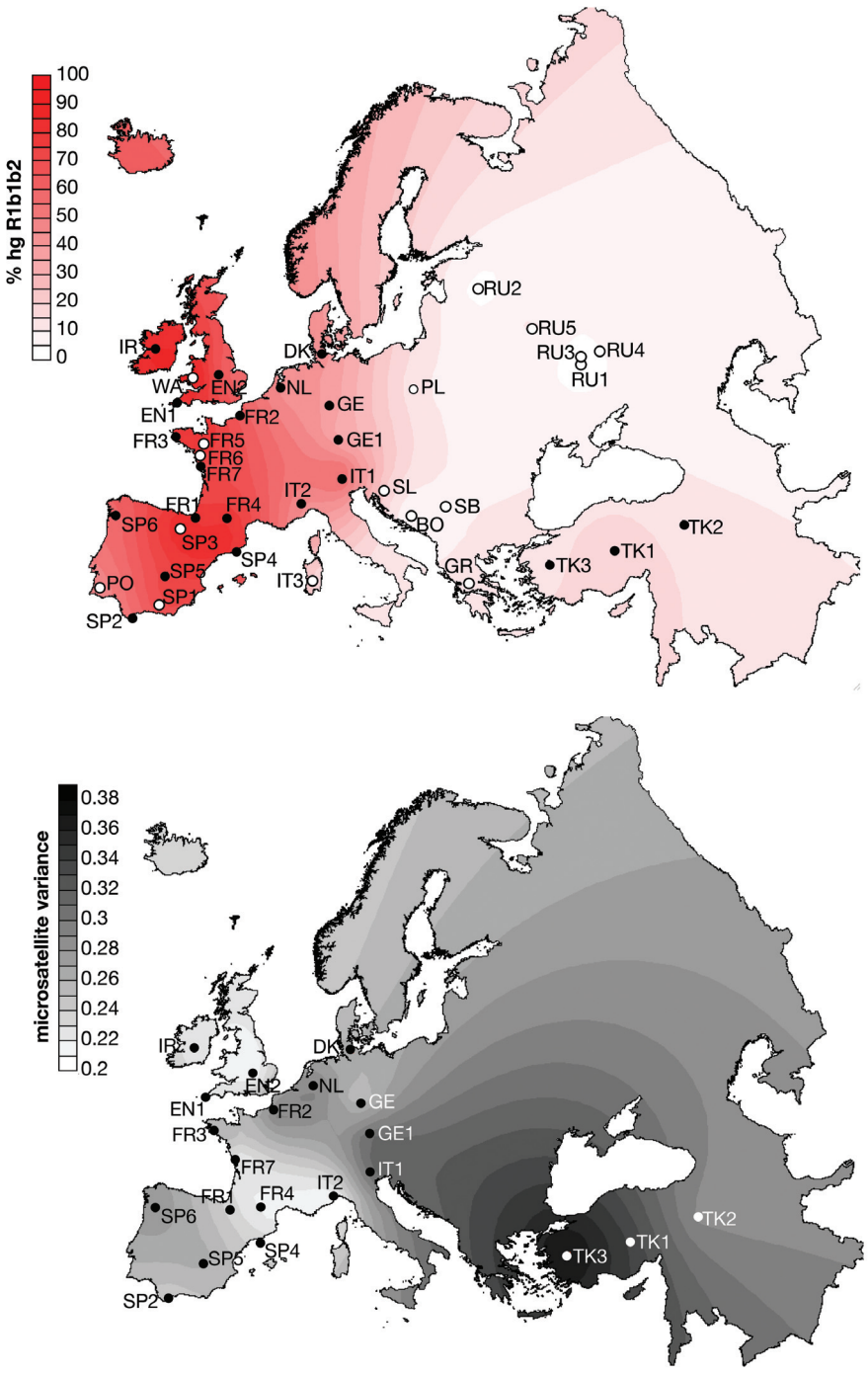


FIGURE 19.9

Clinal gradient in Y chromosome variation across Europe. Top, frequency of haplogroup R1b1b2, one of the most common haplogroups in Europe, showing a clinal increase in frequency from east to west. Bottom, variance in microsatellite (STR) loci associated with this haplogroup, showing a decrease in variance from east to west. These two observations together suggest an origin for this haplogroup in the east (where it has the highest associated microsatellite variance) and spread to the west (where it has the highest frequency but lower microsatellite variance). Reprinted with permission from Balaresque, P., et al., "A predominantly Neolithic origin for European paternal lineages," *PLoS Biology* 8:e1000285, 2010.

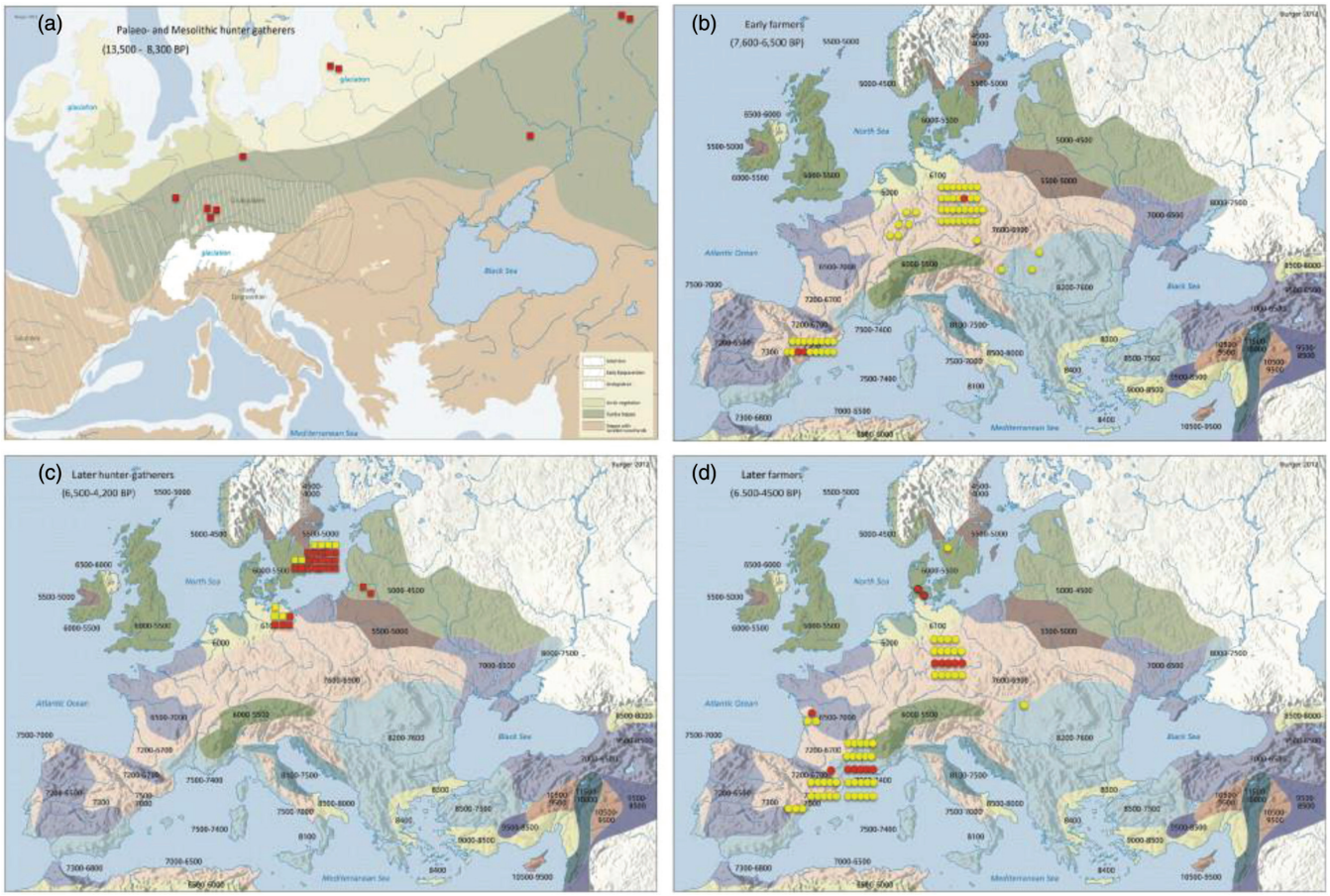


FIGURE 19.10

The mtDNA landscape of prehistoric Europe at four different times, showing mtDNA types recovered from (a) Paleo- and Mesolithic hunter-gatherers; (b) early farmers; (c) later hunter-gatherers; and (d) later farmers. Red, haplogroup U; Yellow, all other haplogroups. Reprinted with permission from Pinhasi, R., et al., "The genetic history of Europeans," *Trends in Genetics* 28:496, 2012.

(although closest to northern Europeans, e.g., Finns and Russians). These results suggest that (1) the hunter-gatherers represent a gene pool that no longer exists as such in Europe; (2) farming spread to Sweden via people who genetically most closely resemble contemporary southern European populations; (3) the contemporary Scandinavian population is not genetically identical to either the hunter-gatherers or the farmer but rather is descended from a mixture of the two. Similar results were then obtained from genome-wide data from skeletons from two 7000-year-old Iberian hunter-gatherers (Sánchez-Quinto et al. 2012): namely, they differ genetically from all contemporary European populations but are more similar to northern Europeans than to southern Europeans.

These results have recently been complemented with a veritable flood of genomic data from literally hundreds of skeletal remains (Allentoft et al. 2015; Haak et al. 2015; Lazaridis et al. 2014; Mathieson et al. 2015), and confirm genetic discontinuities between the pre-Neolithic and Neolithic populations of Europe. The current picture includes a probable Anatolian source for the early European farmers, as well as a later contribution to European ancestry coming from the steppe region (just west of the Ural Mountains in Russia); all modern European populations are thus a composite of three sources of ancestry: European hunter-gatherers, Anatolian farmers, and a steppe population called the Yamnaya. Overall, the genetic results overwhelmingly support the demic diffusion of agriculture to Europe, leading some commentators to proclaim that the problem of the "Neolithization of Europe" has been solved.

However, while practically all agricultural expansions investigated via genetic analyses have shown that such expansions involve demic rather than cultural diffusion, that is not to say that all agricultural expansions were the same. For example, the genetic evidence strongly suggests that the spread of agriculture across sub-Saharan Africa, associated with Bantu-speaking groups, was accompanied by sex-biased admixture (Wood et al. 2005). There is a much

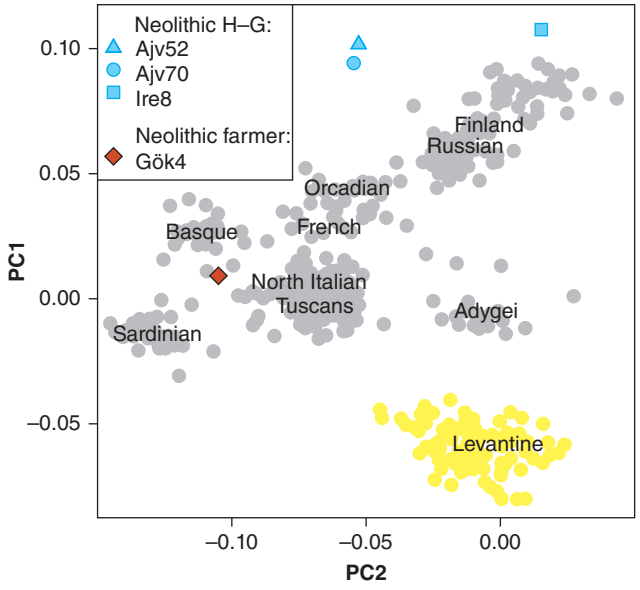


FIGURE 19.11

PC analysis of genome-wide data from three Neolithic hunter-gatherer skeletons (blue) and one Neolithic farmer (red) skeleton, all from Sweden. Note that the farmer groups with southern Europeans, while the hunter-gatherers are distinct from all other groups, although closest to northern Europeans. Reprinted with permission from Skoglund, P., et al., "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe," *Science* 336:466, 2012.

stronger signal of this expansion in the Y chromosomes of Bantu-speaking groups than in the mtDNA, suggesting that as the Bantu-speaking agriculturalists spread across Africa, they admixed heavily with the hunter-gatherer females but not the males. Contrast this with the expansion of Austronesians through Near Oceania on their way to Remote Oceania (discussed in detail in Chapter 16), which involved much more admixture with Papuan males than with Papuan females (probably as a consequence of matrilocality), and you can begin to appreciate the diversity and complexity that underlies simple statements such as "agricultural expansions primarily involved demic diffusion." Clearly, every expansion has its own story to tell, and genetics can help decipher that story.

USING GENETIC ANALYSES TO LEARN MORE ABOUT CULTURAL PRACTICES: LANGUAGE REPLACEMENTS

Agriculture is not the only cultural practice that can spread in principle via either cultural or demic diffusion. Language is another example—indeed, many agricultural expansions were also accompanied by the spread of associated language families, such as the case we've already seen of Austronesian languages probably spreading initially with rice farming from Taiwan, or Bantu languages spreading with the farming of millet and other crops across sub-Saharan Africa. In addition, it does sometimes happen that populations end up speaking languages that are entirely different from their geographic neighbors, and genetic analyses can help understand how this happened. Take, for example, the case of the Caucasus (described briefly in Chapter 10 as an example of AMOVA). As shown in Figure 19.12, most of the languages spoken in the Caucasus belong to either the North Caucasian or South Caucasian language groups. However, there are also some populations in the Caucasus that speak languages belonging to language families found elsewhere, such as Armenian (an Indo-European language) or Azerbaijani (a Turkic language). The question then arises as to how this situation, with groups whose geographic neighbors are not their linguistic neighbors, came to be; in principle, there are two possible explanations (Figure 19.13). If we focus for the moment just on Armenian, while recognizing that exactly the same possibilities hold for Azerbaijani, then one possibility is that a group of people speaking an Indo-European language came to the Caucasus, expanded, and became the current Armenians. The other possibility is that a very small group of people speaking an Indo-European language came to the Caucasus and their language spread to other people who did not originally speak the language. It should be clear that in the first case, Armenians should genetically be more closely related to other Indo-European–speaking groups than to their neighboring groups in the Caucasus, while in the second case, Armenians should genetically be more closely related to their neighboring groups in the Caucasus than to Indo-European–speaking groups elsewhere. In other words, if the Armenian language is spread by a migration of proto-Armenian–speaking people (i.e., speaking a language ancestral to present-day Armenian) to the Caucasus followed by population increase, then Armenians should genetically resemble their linguistic neighbors more than their geographic neighbors. Conversely, if the Armenian language is spread via **language replacement**, then Armenians should genetically resemble their geographic neighbors more than their linguistic neighbors.

Initial studies based on mtDNA strongly favored the language replacement hypothesis for both Armenian and Azerbaijani (Figure 19.14), as speakers of both languages are genetically more similar to their geographic neighbors (other groups in the Caucasus) than they are to their linguistic neighbors (Nasidze and Stoneking 2001). There are a number of ways by which such language replacements can potentially occur, but one favorite hypothesis is called **elite dominance**. According to this model, a relatively small group of

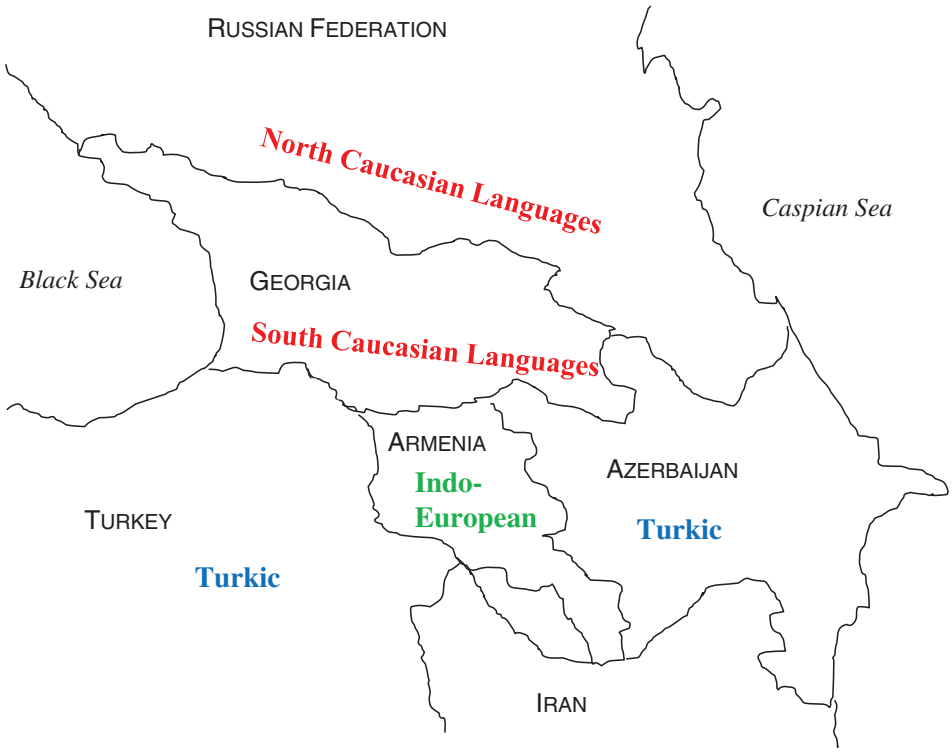


FIGURE 19.12

Language map of the Caucasus.

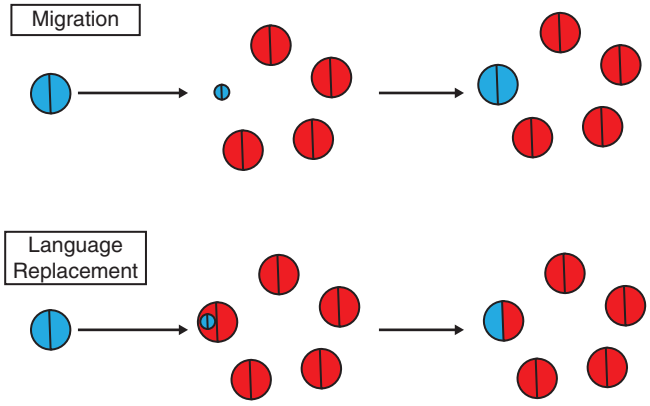


FIGURE 19.13

Migration versus language replacement as potential explanations for how populations that are geographic neighbors speak unrelated languages. Each circle represents a population, with the color of the left half indicating the language family affiliation and the right half the genetic relationships. Top, a small group migrates to a new area and expands, keeping both the language and the genetic relationships of the source group. Bottom, a small group migrates to a new area and the language spreads, but not the genes, resulting in a mismatch between the linguistic and genetic relationships of the affected population.

people move to a new region and either impose their language by fiat or the language is adopted by others because it is seen as socially prestigious. In such cases, one might expect to see a greater contribution of Y chromosomes than mtDNAs from the incoming group, if it was males who were involved either primarily in the migration (say, as conquerors) or in spreading the influence of the elite group (and, concomitantly, spreading their genes!). However, in the specific case of the Armenians and Azerbaijani, NRY variation gives exactly the same picture as mtDNA variation (Nasidze et al. 2003): both groups resemble their geographic neighbors, not their linguistic neighbors (Figure 19.14). Thus, if the Armenian and Azerbaijani languages do reflect language replacements via elite dominance (and many linguists suspect this is the case, based on linguistic evidence), then the incoming group had a negligible genetic impact compared to their linguistic impact. In any event, investigating potential replacements and other language contact situations is a very fruitful area for genetic research.

USING GENETIC ANALYSES TO LEARN MORE ABOUT CULTURAL PRACTICES: DATING THE ORIGIN OF CLOTHING

The final example of how genetic analyses can provide insights into cultural practices involves a molecular

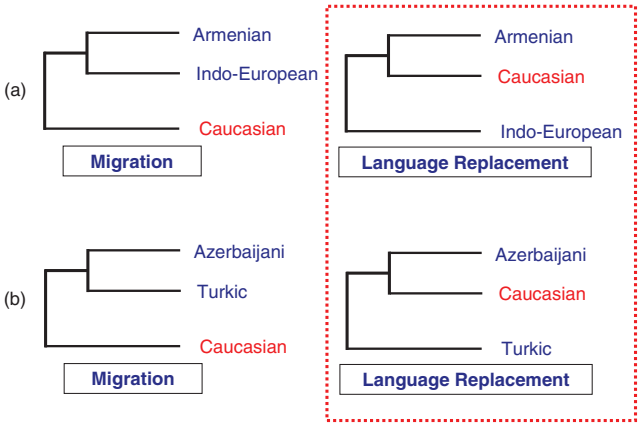


FIGURE 19.14

Genetic relationships expected for Armenian and Azerbaijani populations compared to their linguistic neighbors (Indo-European and Turkic-speaking groups, respectively) and their geographic neighbors (other groups in the Caucasus) under two different hypotheses concerning their origins. (a) Left, relationship expected for Armenians under the migration hypothesis; right, relationship expected for Armenians under the language replacement hypothesis. (b) Left, relationship expected for Azerbaijani under the migration hypothesis; right, relationship expected for Azerbaijani under the language replacement hypothesis. The dotted box indicates that the genetic evidence supports the language replacement hypothesis for both Armenians and Azerbaijani.

genetic approach to dating the origin of clothing. Here, though, instead of analyzing genetic variation in humans, the insights come from analyzing genetic variation in one of our parasites—namely, lice. I first got interested in lice several years ago, when my oldest son came home from school one day with a flyer stating that one of his classmates had lice, and what signs we should be looking for in our own children in case they also had lice. Among the various "fun facts" about lice in this brochure, two statements that caught my eye were that lice parasitize only humans (so they can't live on your dog or cat or other pet), and they can't survive away from the human body more than about 24 hours (so they can't live for a long period of time just in your bedding, carpeting, furniture, etc.). In scientific terms, then, lice are obligate parasites of humans—no other host will do—and, therefore, the spread of lice around the world must have occurred via the spread of humans around the world. Thus, by studying genetic variation in lice, we might gain more insights into human migrations and dispersal—one of my favorite topics for research.

I filed this information away until I had the opportunity to follow up on it, and then when I began reading about lice, I learned that they were potentially even more interesting than I had thought. It turns out that many creatures have their own species of lice—mammals have lice, birds have lice, even fish have lice. However, humans are pretty special in that while most creatures have just one kind of lice, we have three different kinds of lice (so, keep that fact in mind in case you are ever asked how humans differ from other organisms!). Two of these are the head louse, *Pediculus humanus capitis*, and the body louse, *Pediculus humanus humanus* (we'll get to the third kind of louse later on). These are so closely related that it is difficult to distinguish them visually (Figure 19.15)—indeed, the main difference between them is in their ecology. The head louse, as the name suggests, lives and feeds exclusively on the human scalp, whereas the body louse feeds on the human body, but actually lives in clothing, where it lays its eggs. And if you ask yourself how this difference would arise, the answer that seems logical is that before we had clothing, we had only one type of lice, namely, head lice. But then when clothing was invented, it became available as a new ecological niche; head lice moved into the clothing and then adapted to this new "environment," evolving to become body lice. And if that is indeed the case, then the divergence between head and body lice would have occurred when clothing became important in human evolution, which we can then date by using a molecular clock approach to

FIGURE 19.15

Lovely lice. From left to right: head louse, body louse, pubic louse, and chimpanzee louse. Head louse and body louse, author; pubic louse and chimpanzee louse, modified with permission from Reed, D.L., et al., "Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice," *BMC Biology* 5:7, 2007.

date the divergence between head and body lice. So, we did just that (Kittler et al. 2003).

So to do this, we first needed samples of lice from around the world, which were not so easy to come by. We contacted various clinics, hospitals, prisons, and military bases—sometimes receiving outraged responses as to how dare we suggest that there might be lice at their facility—and eventually managed to amass a fairly diverse sample of head and body lice from around the world. Recall from Chapter 12 that to use a molecular clock approach, we need a calibration point to get an estimate of the mutation rate (i.e., we need to know how fast the clock is ticking). We therefore obtained samples of chimpanzee lice (*Pediculus schaeffi*) from a sanctuary in Uganda and assumed that the divergence of human from chimpanzee lice would have occurred when their hosts, namely, humans and chimpanzees, diverged (cospeciation of parasites and their hosts is indeed usually the case, although exceptions are known).

The next step was to obtain mtDNA sequences from the samples and analyze them. Nucleotide diversity values were significantly bigger for head lice (3.4%) than for body lice (0.2%), which is reassuring as it fits with the assumption that body lice originated from head lice (and that this origin was accompanied by a reduction in population size, meaning that body lice genetic diversity is a subset of head lice genetic diversity). Nucleotide diversity values were also significantly bigger for African lice (3.3%) than for non-African lice (1.7%), suggesting an African origin for lice, which is also reassuring since the genetic evidence points toward an African origin for their host—namely, us! A neighbor-joining tree of the mtDNA sequences (Figure 19.16), rooted with the chimpanzee louse sequence, indicates that the deepest splits in the tree all involve sequences from head lice. Body lice mtDNA sequences are all in a clade with head lice sequences (with one sequence type actually shared by head and body lice) that has many sequences branching off simultaneously, suggestive of a population expansion. The fact that body lice sequences do not form their own separate clade but rather fall in (and are even shared with) head lice sequences is an indication of a recent origin of body lice. Alternatively, this could be explained by ongoing gene flow (hybridization) between head and body lice, but this does not seem likely as studies of people who were unfortunate enough to be infested with both head and body lice have shown that the head and body lice from the same person are genetically more different than are head lice from different people or body lice from different people (Leo et al. 2005)—if head lice and body lice from the same person were

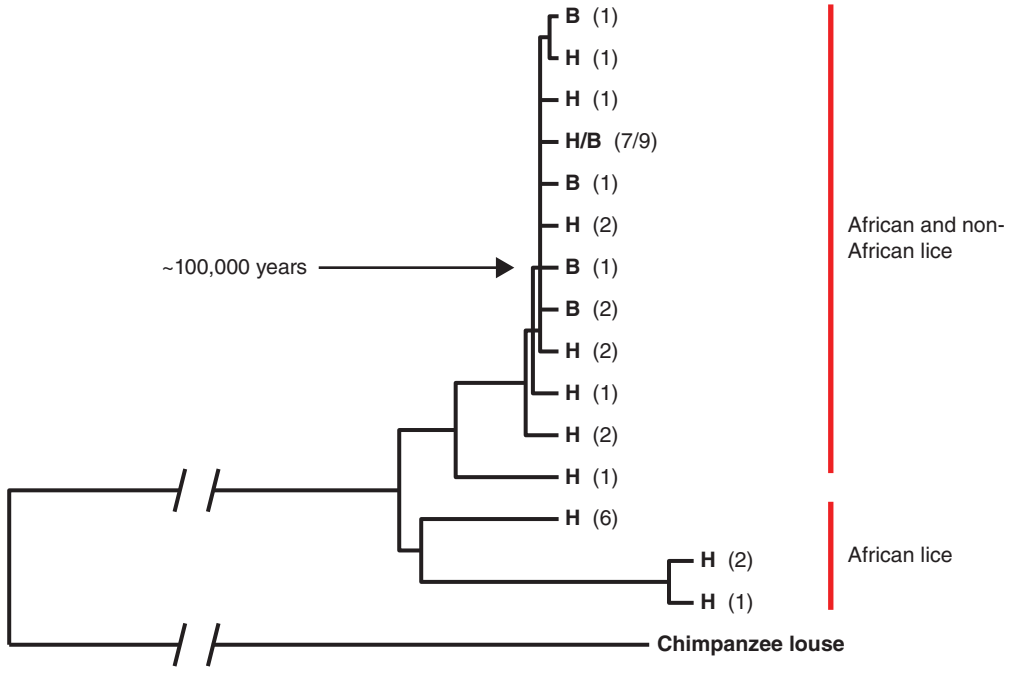


FIGURE 19.16

Neighbor-joining tree for lice mtDNA sequences. H, head lice; B, body lice; numbers within parentheses indicate number of lice with that sequence. The age of the clade containing all of the body lice sequences is indicated. Modified with permission from Kittler, R., et al., "Molecular evolution of Pediculus humanus and the origin of clothing," *Current Biology* 13:1414, 2003.

interbreeding, then they should be genetically more similar.

By assuming that the genetic divergence between the human and chimpanzee lice happened when their human and chimpanzee hosts speciated, namely, about 5.5 million years ago, then the molecular clock approach dates the origin of the clade containing the body lice mtDNA sequences to about 100,000 ± 40,000 years ago, which then provides an upper limit to the divergence between head and body lice (because of ancestral polymorphism, as discussed in Chapter 12)—and, by inference, is when clothing was used widely enough to serve as a new niche for the ancestors of body lice. While this is a rather substantial time range (because of the inevitable large variances associated with molecular clock dating, as discussed in Chapter 12), it is noteworthy that it does overlap the estimated time for the first dispersal of modern humans from Africa. Moreover, there is a signature in the lice mtDNA sequences of a population expansion at this time—just as there is a signal of population expansion (following a bottleneck) in modern humans associated with the first migrations out of Africa. The admittedly highly speculative scenario that these results suggest is that the invention of clothing may have been a contributing factor in the expansion of modern humans out of Africa into more extreme latitudes. Most likely, the earliest clothing was some form of animal hide or skin, which because of its similarity to human hair facilitated the movement of lice into this new niche.

How do the genetic results compare to fossil evidence for the origin of clothing? Clothing alas does not fossilize, so we have to rely on indirect evidence. The earliest stone tools that are unambiguously associated with clothing (needles and the like) are at most about 40,000 years old (although there have been suggestions that some pierced shell beads that are upward of 100,000 years old may have been used as buttons). To be sure, generalized scraping tools that could have been used to prepare hides for clothing go back hundreds of thousands of years, but we don't know whether that's what they were used for, or whether they were used for other purposes. In sum, the available archaeological evidence is in keeping with a relatively recent origin of clothing (i.e., within the past 100,000 years or so).

Any time you say something about what modern humans were up to, sooner or later somebody wants to know whether Neandertals were also doing the same thing. So, what about Neandertals (and Denisovans and other archaic humans)? Did they have clothing? Figure 19.17 shows the evolutionary relationships of Neandertals and modern humans, along with the presumed origin of clothing. Since the evidence from lice indicates that modern humans invented clothing well after their divergence from Neandertals, it then

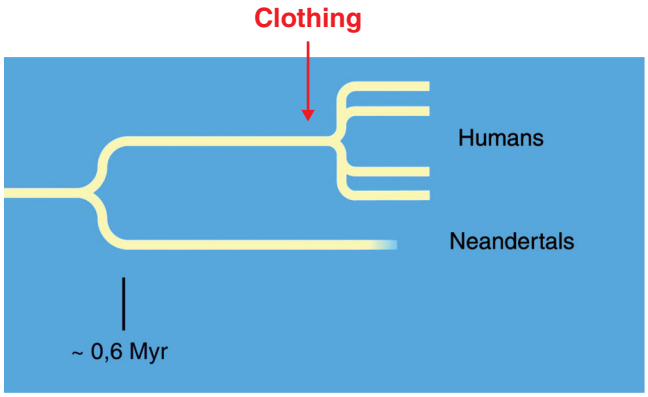


FIGURE 19.17

Evolutionary relationships of humans and Neandertals, along with the presumed origin of clothing.

follows that Neandertals did not "inherit" clothing from their common ancestor with modern humans. So, either Neandertals invented clothing independently from modern humans or they did not have clothing. Now, there is nothing in what we can infer about the cognitive abilities of Neandertals from the available fossil (and genetic) evidence that would indicate that they were too stupid to come up with the idea of clothing. As far as we can tell, they would have been perfectly capable of coming up with clothing on their own. Moreover, it is also certainly the case that reconstructions of Neandertals invariably show them as having clothing (e.g., see Figure 19.18, left), probably because they were living in a rather cold climate. But it is possible that Neandertals did not have clothing, even in the cold European climate, if they still had body hair (e.g., see Figure 19.18, right). That is, one of the prominent differences between us and all other primates (indeed, nearly all other mammals) is that we have lost most of our body hair. When this happened during human evolution is still a matter of conjecture, but one possibility is that it happened relatively recently, after our ancestors diverged from Neandertals, and so Neandertals still had body hair—and thus, no need for clothing. So if we could figure out when body hair was lost during human evolution, then we could get a better idea as to whether or not Neandertals still had body hair.

How could we determine when body hair was lost? Well, we could get very lucky and find a Neandertal frozen in ice and then directly see how much body hair was present, but so far that hasn't happened. Or, if we knew the genetic mutation(s) responsible for loss of body hair, we might be able to date the mutation(s) via molecular clock approaches—but so far, that hasn't happened either. Instead, our best evidence for when humans lost body hair comes from the third kind of louse that parasitizes humans—namely, the

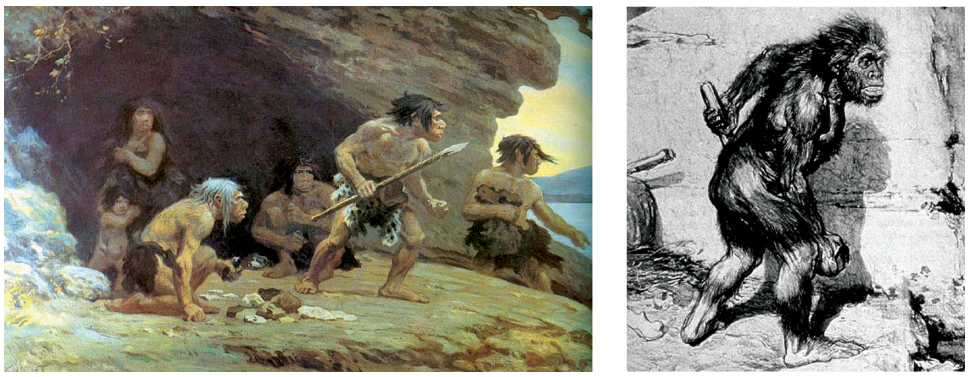


FIGURE 19.18

Two hypothetical reconstructions of Neandertals, one with clothing and without body hair, and one without clothing but with body hair. Left, reprinted with permission from Wikimedia Commons (http://commons.wikimedia.org/wiki/File:Le_Moustier.jpg); Right, see Figure 14.8.

pubic louse (*Pthirus pubis*), also known as the crab louse (Figure 19.15). As the name suggests, the pubic louse lives and feeds exclusively on the pubic region. And if you ask yourself how did this come to be, the answer that seems to make the most sense is, well, before our ancestors lost body hair, they were parasitized by just one kind of lice all over their body—after all, this is the case for all other mammals. But then with the loss of body hair, lice become "geographically" isolated in the pubic region and the head, and then classic "allopatric" speciation ensued due to the geographic barrier (viz., the hairless torso), leading to the formation of pubic lice and head lice as different species. And if this was indeed the case, then by dating the divergence of pubic and head lice with a molecular clock approach, we can get an estimate as to when our ancestors lost their body hair.

Sounds great, right? Unfortunately, it's not quite that straightforward, as there is one small complication. Figure 19.19 outlines the scenario that I have been describing for lice evolution, starting with the common ancestor of humans, chimpanzees, and gorillas (for reasons that will soon be apparent). After the gorilla lineage branches off, the lice that accompanied them evolved into a distinct species; the next event is the divergence between humans and chimpanzees and the corresponding evolution of human and chimpanzee lice. Then, the ancestors of humans lose body hair, leading to the formation of pubic lice and head lice, and finally clothing is invented, leading to the origin of body lice. Looks nice and straightforward, right—except for one tiny fact: namely, the taxonomy of lice does not agree with this scenario! Note in Figure 19.19 that the chimpanzee, head, and body lice are all classified in the same genus, *Pediculus*. The gorilla louse is classified in a different genus, *Pthirus*—which also just happens to be the same genus as the pubic louse!

Before jumping to any hasty conclusions, it should be noted that maybe this taxonomic classification is incorrect, and the pubic louse really belongs in the same genus as the other human lice. It would not be the first time that a taxonomic classification of a group of organisms does not coincide with their evolutionary relationships (as determined via DNA analyses). So, clearly the thing to do is to get DNA from all of these lice (including gorilla lice) and see what the DNA has to say. David Reed, a mammalogist who studies mammals, their parasites, and even parasites of the parasites (e.g., bacteria in lice), did just that (Reed et al. 2007), and the results (Figure 19.20) do confirm that the taxonomists actually got it right: namely, the human pubic

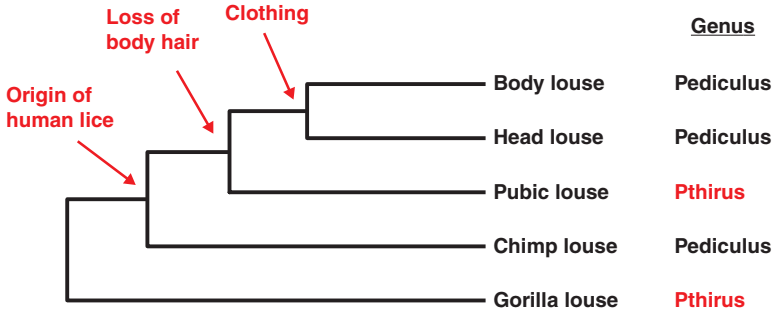


FIGURE 19.19

An idealized view of lice evolution, with important events noted, and with the genus for each louse species indicated. Note that the classification of the human pubic louse and the gorilla louse as belonging to the same genus (*Pthirus*) contradicts this view of lice evolution.

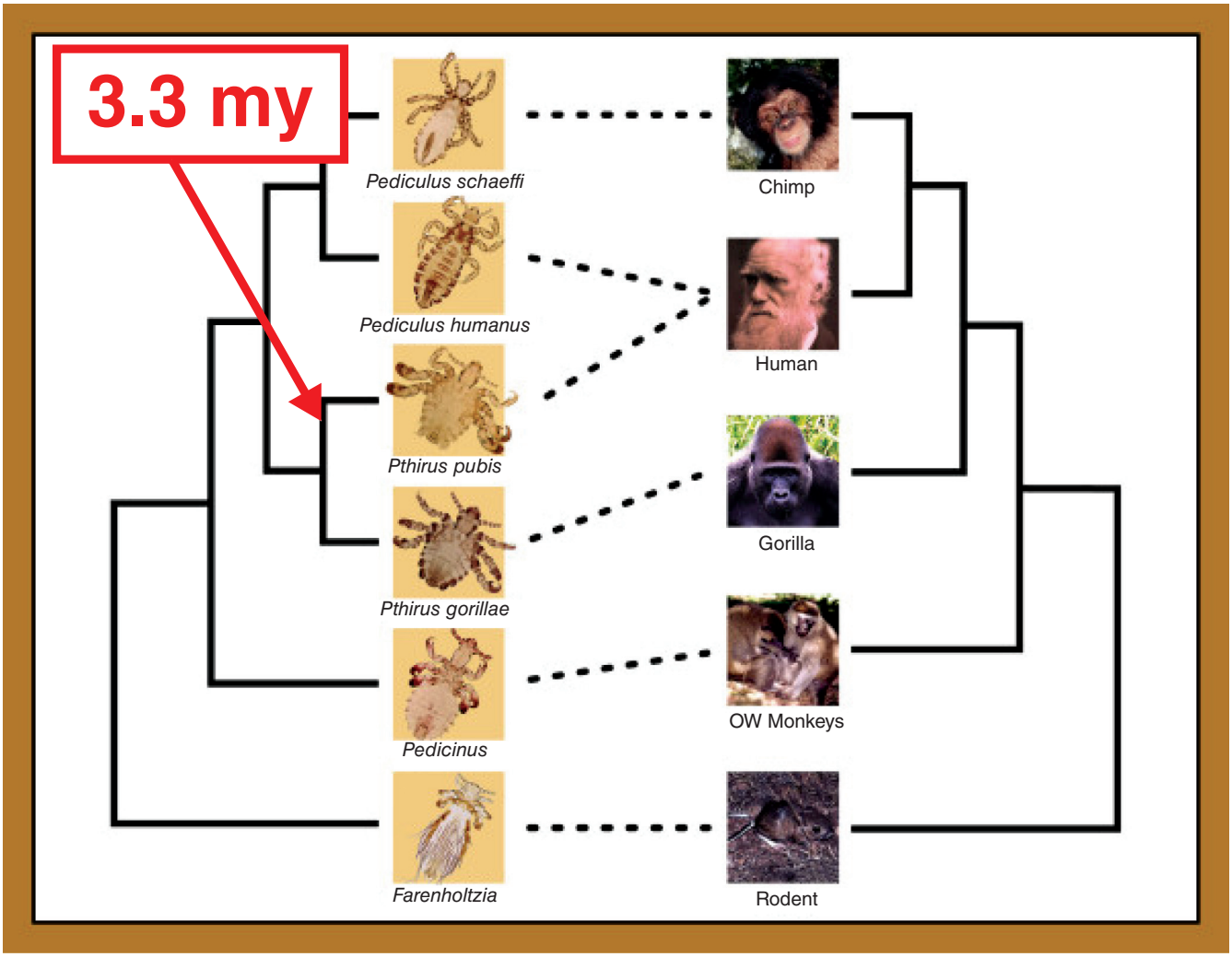


FIGURE 19.20

Results from DNA sequence comparisons confirm that the taxonomic classification of lice does indeed reflect their evolutionary relationships. In particular, the human pubic louse is more closely related to the gorilla louse than to other human lice, with a divergence time of about 3.3 million years ago. Reprinted with permission from Reed, D.L., et al., "Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice," *BMC Biology* 5:7, 2007.

louse is indeed more closely related to the gorilla louse than it is to the other human lice. I will refrain from speculating how it happened that humans got pubic lice from gorillas, but it does seem as if our ancestors have some explaining to do!

Anyway, what do these results tell us about the loss of body hair? Using a molecular clock approach, the estimated divergence time between the gorilla louse and the pubic louse is about 3.3 million years. It seems reasonable to suppose that humans had lost body hair by the time this happened, so pubic hair would be available as a new niche for the gorilla lice to colonize and then adapt to, evolving to become pubic lice. This then suggests that loss of body hair happened relatively early in human evolution, and that our ancestors went around for a long time without either body hair or much in the way of clothing—presumably because they were living in a warm environment (i.e., Africa). Moreover, this time of 3.3 million years ago for loss of body hair fits nicely with the hypothesis that losing body hair was a thermoregulatory adaptation that promoted cooling via sweat evaporation, necessitated by our ancestors moving into a hot and arid Savannah environment not long after diverging from the chimpanzee lineage (Jablonski 2006). Of course, the loss of body hair relatively soon after our lineage diverged from that of chimpanzees is compatible with other hypotheses for the loss of body hair, such as sexual selection or to help avoid parasites (which would be rather ironic since loss of body hair apparently enabled

gorilla lice to parasitize us!). It is also worth pointing out that this relatively old time for the loss of body hair would indicate that Neandertals (and Denisovans and all other members of our genus, *Homo*) were also hairless, and thus those archaic humans who lived outside Africa probably invented clothing independently (and thus, perhaps even developed their own species of body lice?). In any event, the take-home message is that analyzing genetic variation in these interesting parasites has led to some novel insights into cultural practices, such as when we began making frequent use of clothing, as well as into aspects of human evolution that cannot be directly observed in the fossil record, such as the loss of body hair.

CONCLUDING REMARKS

We began this chapter by asking whether or not humans are still evolving, because of the influence of culture, and the clear message is that humans have been evolving and are continuing to evolve, and we are doing so not just despite culture but because of culture. For those of you who are still skeptical about the prospects for any "meaningful" biological evolution when it comes to humans, consider the following. Recent sequencing studies of the entire genome from families (e.g., Roach et al. 2010) have revealed that every child has on the order of 50–100 new mutations—that is, new genetic variants not present in either parent. There are something like 130 million children born each year, so if you do the math, there are about 6.5–13 billion new mutations coming into this world every year, which is enough for each nucleotide in the human genome to have mutated on average about 2–4 times. So, there is lot of opportunity for new, selectively advantageous mutations to arise. The take-home message: if something happens that has an effect on our biological fitness, and we are either unwilling or unable to deal with it via cultural means (think of diseases such as malaria and AIDs that still plague us), then rest assured that we will evolve via natural selection and adapt to the changed circumstances—as we have in the case of malaria, as discussed in Chapter 5, and may be doing so in the case of AIDs, with some mutations known that decrease susceptibility to AIDs (Liu et al. 1996). We've done so time and time again during the course of our evolutionary history, and all indications are that we will continue to do so—unless, of course, the change in circumstances is too drastic for us to respond to either biologically or culturally, for example, a massive asteroid impact, catastrophic climate change, global nuclear warfare, or something equally unpleasant to contemplate. In which case, we will suffer the same fate as the vast majority of species that have ever existed on this planet—namely, extinction. And on that cheery note, we'll end this chapter and turn to the future of molecular anthropology (assuming there is one).

CHAPTER 20

ONGOING AND FUTURE DEVELOPMENTS IN MOLECULAR ANTHROPOLOGY

In this final chapter, we will consider some of the ongoing and (likely) future developments in molecular anthropology. This is a time of great change, both in terms of methods for producing genetic data (e.g., there is already talk of "third-generation" sequencing technologies that will render the current "next-generation" platforms obsolete) and in what we can learn by analyzing such data. At least some of what is in this book is likely to be out of date by the time you read it, so in an attempt to atone for such shortcomings, in this chapter, we will discuss some of the areas that are likely to be changing the fastest, and what future developments are likely to bring. The wide-ranging topics we will touch on include more genetic data as well as variation in other molecules (the other "omics"), more (and different) analyses, relating phenotypes and genotypes (with skin pigmentation variation discussed in detail as an example), and finally a look at what personal ancestry and genomics testing means for molecular anthropology. At the same time, this chapter will (hopefully) reinforce some of the main points made previously.

MORE—AND DIFFERENT KINDS OF—DATA: THE OTHER "OMICS"

A very safe prediction that we can make is that we will continue to see more and more genome-wide data. As we have seen, SNP chips (even with the drawback of ascertainment bias) are providing important new insights into the genetic history of human populations, particularly (but not only) with respect to identifying previously unsuspected migrations and admixture events. The costs of such chips are rapidly dropping, and new versions are coming out all the time, including versions that try to minimize the effect of ascertainment bias as well as versions targeted at individuals interested in their own ancestry—more on the increasing interest in "personal genomics" and what this means for molecular anthropology later in this chapter.

But the real revolution that is coming in terms of genome-wide data is the growing availability of partial and "complete" genome sequences, thanks to the advances in next-generation sequencing platforms. For the uniparental markers, it is fast becoming routine to generate complete mtDNA genomes and partial Y chromosome sequences, thereby maximizing the information that one can get concerning the maternal versus paternal history of human populations. As for the rest of the genome, at the current cost of a few thousand dollars per sequence, it is still too costly to carry out genomic sequencing at the population level, that is, from several individuals from each of several populations (except in the context of large consortia projects such as the 1000Genomes project). However, with so-called "third generation" sequencing platforms on the horizon, within a few years it is quite likely that complete genome sequencing will be the method of choice for investigating population history.

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

And in the meantime, there is a lot that can be learned about population history, even from a single genome sequence—recall from Chapter 12 that we can, for example, infer the history of population size change from just a single sequence. Moreover, partial genome sequences (such as exome sequencing) can already produce more data than we know what to do with (more on this later in this chapter).

There is also much promise for technological advances in recovering and analyzing ancient DNA. Just in the few years that next-generation sequencing has become available, we've seen genome sequences from archaic humans (Neandertals and Denisovans) as well as from ancient modern humans. And just in the last year or so, further technical developments have enabled the recovery of high-quality DNA sequences from archaic human remains that are every bit as good as those from contemporary samples. Moreover, genome-wide data are now available from hundreds of remains that range in age up to 10,000 years or so, with lots more on the horizon, and these studies are providing some very interesting and important insights into human population history. So, it is a pretty safe bet that we can expect to see lots more in the way of ancient DNA studies. Still, there are many interesting parts of the world where the authentic DNA either is not present in sufficient quantities or is too highly fragmented to be of any use (in particular, where it is hot and humid), so barring some technical breakthrough (which one should never bet against!) there are some parts of the world where we will have to rely on what we can infer from studying genetic variation only in contemporary populations.

In this book, we have defined molecular anthropology as the study of one type of molecule, namely, DNA or the genome, to address questions of anthropological interest. But there is more to life than just DNA—there are other types of molecules that can be studied in addition to genomics, and this is where the other "omics" come in. **Transcriptomics** is the study of the transcriptome, or RNA, that is produced from DNA. We've already seen evidence that changes in the regulation of gene expression are perhaps even more important evolutionarily than changes in the structure of genes, so you might think that studying variation in the abundance of various RNA transcripts in different species/populations would provide insights into how gene regulation has evolved. And to be sure, there have been some intriguing results. For example, when one compares changes in gene expression in different tissues within and between humans and chimpanzees, what stands out is that there has been a large acceleration in the rate of transcriptome evolution along the human lineage in brain tissue relative to other tissues (Figure 20.1). This would suggest that there has been more selection along the human lineage for changes in gene expression in the brain than in other tissues, and hence changes in brain function were important in human evolution—the latter may seem rather obvious when you compare humans to chimpanzees, but

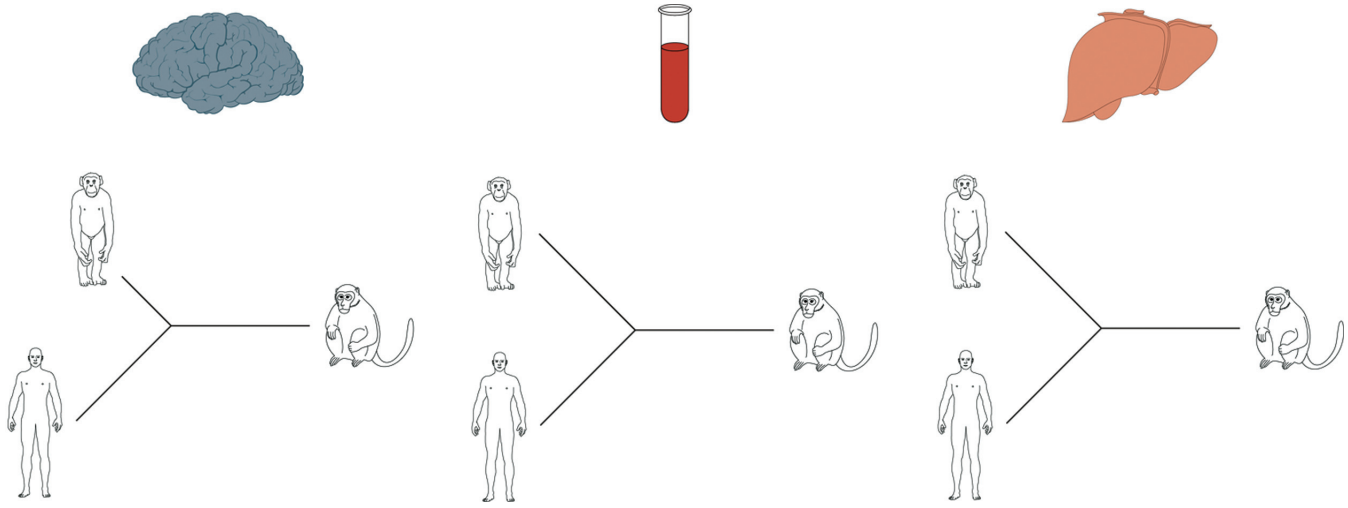


FIGURE 20.1

Relative amounts of changes in gene expression in brain, blood, and liver from humans, chimpanzees, and rhesus macaques. Genes expressed in brain tissue (left) have changed much more along the human lineage than have genes expressed in blood (center) or liver (right), consistent with the view that genetic changes influencing brain function and development were more important during human evolution than genetic changes influencing blood or liver function and development. The information on which this figure is based is from Enard, W., et al., "Intra- and interspecific variation in primate gene expression patterns," *Science* 296:340, 2002.

seeing this reflected in transcriptome evolution provides yet more evidence for the evolutionary importance of changes in gene regulation.

There have also been a few studies that examined variation in gene expression among individuals from different populations (e.g., Stranger et al. 2007; Zhang et al. 2008). In addition to identifying differentially expressed genes, such studies have also identified so-called **eQTLs** (expression quantitative trait loci), which are DNA polymorphisms that are associated with differences between individuals in the amount of RNA that is transcribed from a particular gene. But transcriptomic studies are still a long way from identifying the important changes in gene regulation between populations/species, for several reasons. There is evidence that, like the genome, the transcriptome evolves mostly according to a neutral model (Khaitovich et al. 2004), in that the transcriptome seems to change at a constant rate over time (Figure 20.2). So, just like with the genome, when we compare transcriptome differences between humans and chimpanzees, or between different human populations, the challenge becomes one of distinguishing neutral changes from selected changes, with the former expected to vastly outnumber the latter.

Moreover, current transcriptome studies are limited in terms of the tissues and/or developmental stages that can be studied—the results in Figure 20.1, for example, are all from brains of adult individuals (who all died at various ages from various causes, which is another problem when dealing with transcriptome studies—whether you die in your sleep or in a car accident will influence what RNA your brain is making at the time of death). However, at least some of the evolutionarily important changes in brain gene expression that differentiate humans from chimpanzees may very well be differences in gene expression that occur during fetal brain development, and hence will not show up in studies of adult brains. But ethically there's no way that studies of human or chimpanzee fetal brain tissue can be carried out.

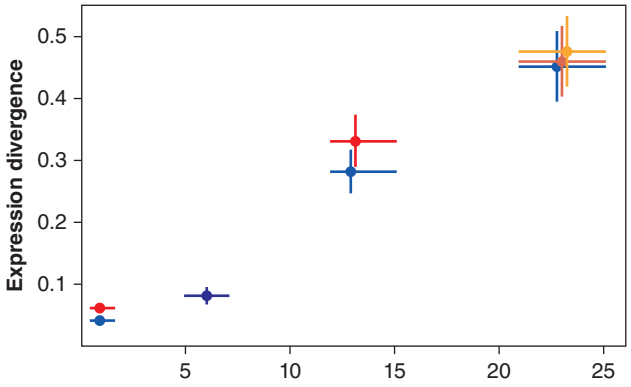


FIGURE 20.2

Amount of divergence in gene expression in brain tissue versus divergence time (X axis, in millions of years) for humans and various nonhuman primates. Colors indicate species compared (orange with humans, blue with chimpanzees, purple between humans and chimpanzees, yellow is between orangutan and rhesus macaque). Modified with permission from Khaitovich, P., et al., "A neutral model of transcriptome evolution," *PLoS Biology* 5:e132, 2004.

Similarly, our current knowledge of transcriptome differences between different human populations is almost entirely limited to laboratory-cultured lymphoblastoid (blood cell) cell lines, so there is lot of potentially significant variation in gene expression among different tissues that isn't picked up by such studies—for example, the variation in lactase expression between lactose-tolerant and lactose-intolerant individuals isn't detected when studying the transcriptome of lymphoblastoid cell lines. But there is a promising new development, involving a type of cell called iPSCs (induced pluripotent stem cells, discussed in more detail later) that may help alleviate some of the drawbacks of current transcriptome studies.

Recall from Chapter 2 that RNA is the intermediate step in going from a gene to the gene's product, which is a protein. So, we could also think about studying variation in the abundance of different proteins among different populations or species, and the study of protein abundance is known as **proteomics**. In principle, proteomics should get us even closer to understanding variation at the phenotypic level between species or populations, because it is ultimately through the proteins that variation at the genetic level is manifested as phenotypic variation. And, there isn't necessarily a one-to-one correspondence between variation at the level of the transcriptome (i.e., how much RNA is made) and that at the level of the proteome (i.e., how much of the corresponding protein is made)—one individual might make a lot of RNA transcript from a particular gene but it gets degraded quickly, while another individual might make less of that RNA transcript but it is more stable, and the end result might be the same amount of protein produced by that gene in those two individuals despite differences in RNA abundance. However, proteomics is extraordinarily complicated, as you have to be able to identify which of the several thousand different proteins in a tissue is which and then measure how much of each protein is present. Determining the amino acid sequence of a protein is a lot more complicated than determining a DNA sequence—in fact, if you just want to know the amino acid sequence of one particular protein, it's a lot quicker and easier to determine the DNA sequence coding for that protein and then infer the amino acid sequence from the DNA sequence. Moreover,

proteins are subject to all sorts of posttranslational modifications, such as glycosylation (addition of sugar residues), phosphorylation (addition of phosphate groups), and so forth, so you need to take all of this variation into account as well when analyzing proteins. Still, without going into the technical details, there have been some recent promising developments in proteomics that should enable cross-species and cross-population comparisons in protein abundance (i.e., evolutionary proteomics), so that is something to look forward to in the near future.

In addition to DNA, RNA, and proteins, there are other substances present in our bodies, derived mostly from the food that we ingest, and we can study variation in their abundance—this is known as the **metabolome** (because these substances are a by-product of metabolism). Studies of the metabolome are still in their infancy, as there is as yet no convenient method for screening the entire range of metabolites, so currently only a fraction are assayed. Still, there is great interest in screening metabolites to look for correlations with various indicators of health and disease; several countries have started large **biobank** projects in which blood samples are being taken from thousands of individuals (and sometimes with repeated sampling over many years), along with detailed medical histories. The goal is to look for associations between the metabolome and the genome and how these interact to influence human health. And, as has been the case for many of the methods that have been used in anthropological studies, once the medical field figures out a good way to screen the metabolome, we can then expect to see anthropological studies of cross-population and cross-species variation in the metabolome.

BEYOND "YOU": THE MICROBIOME

In this book, we have focused almost exclusively on what we can learn from DNA from our own cells. But there is a lot more to the human body than human cells—in addition to the trillion or so cells that make up your body, you have about 10 trillion bacteria living in and on you. So, what you think of as "you" is only about 10% or so human; the rest is bacteria. The microbial component of the human body is known as the human **microbiome**, and studying the composition of the microbiome is known as **metagenomics**. From an anthropological perspective, the human microbiome has two important implications. First, the spread of human-associated bacteria around the world was accomplished mostly by the spread of humans, so studying variation in such bacteria can be another source of insights into human migrations. To be sure, bacteria are taken up from the environment as well, but at least some kinds of bacteria live in close association with their human hosts and are transmitted mostly between family members or other individuals coming into intimate contact with one another.

We've already seen one example in which studies of a human parasite, namely, lice, provided novel insights into some aspects of human evolution; from the microbial world, the best example comes from a bacteria called *Helicobacter pylori* (or *H. pylori* for short). It turns out that *H. pylori* is the primary cause of stomach ulcers, but it took a long time for the idea that stomach ulcers were a result of an infection to become accepted. It used to be a known medical "fact" that ulcers were caused by stress, smoking, and/or other "lifestyle" factors. Bacteria could not be involved because, as everybody knew, the stomach is too acidic for bacteria to survive. However, as Sherlock Holmes said, "There is nothing more deceptive than an obvious fact," and with improvements in methods for obtaining stomach biopsies, beginning in the late 1970s, the Australian pathologist Robin Warren started coming across a new type of bacteria in biopsies from people with stomach ulcers. Most of his colleagues were skeptical about his findings, attributing them to contamination during the biopsy procedure or to some secondary infection unrelated to ulcers, but Warren convinced a gastroenterologist, Barry Marshall, to collaborate with him, and soon they amassed enough evidence to convince themselves (but not their critics) that there was a strong correlation between the presence of the bacteria and stomach ulcers. However, attempts to culture the new bacteria were fruitless, until culture plates that had been left by accident to incubate over a long holiday weekend showed signs of the new bacteria (previously, the lab techs had been following standard procedures in discarding plates that showed no evidence of bacterial growth after 48 hours). Thanks to this serendipitous accident, they could now grow and study the bacteria in culture, which led to improved methods for diagnosing *H. pylori* and to the first attempts at treating patients with ulcers with antibacterial drugs, which were successful. Still, the medical establishment remained unconvinced that ulcers could be caused by bacteria, as there was one key requirement for causation that had not been fulfilled, namely, treating a healthy host with the bacteria should result in the disease. Marshall and Warren tried infecting pigs, but with no success, so Marshall finally took the drastic step of drinking some culture containing *H. pylori*. When he then developed the symptoms of gastric ulcer a few days later, and a stomach biopsy revealed an active *H. pylori* infection, even the skeptics were convinced. It is now routine to treat stomach ulcers with antibiotics, and for their perseverance in the face of "known facts," Marshall and Warren were awarded a Nobel Prize in

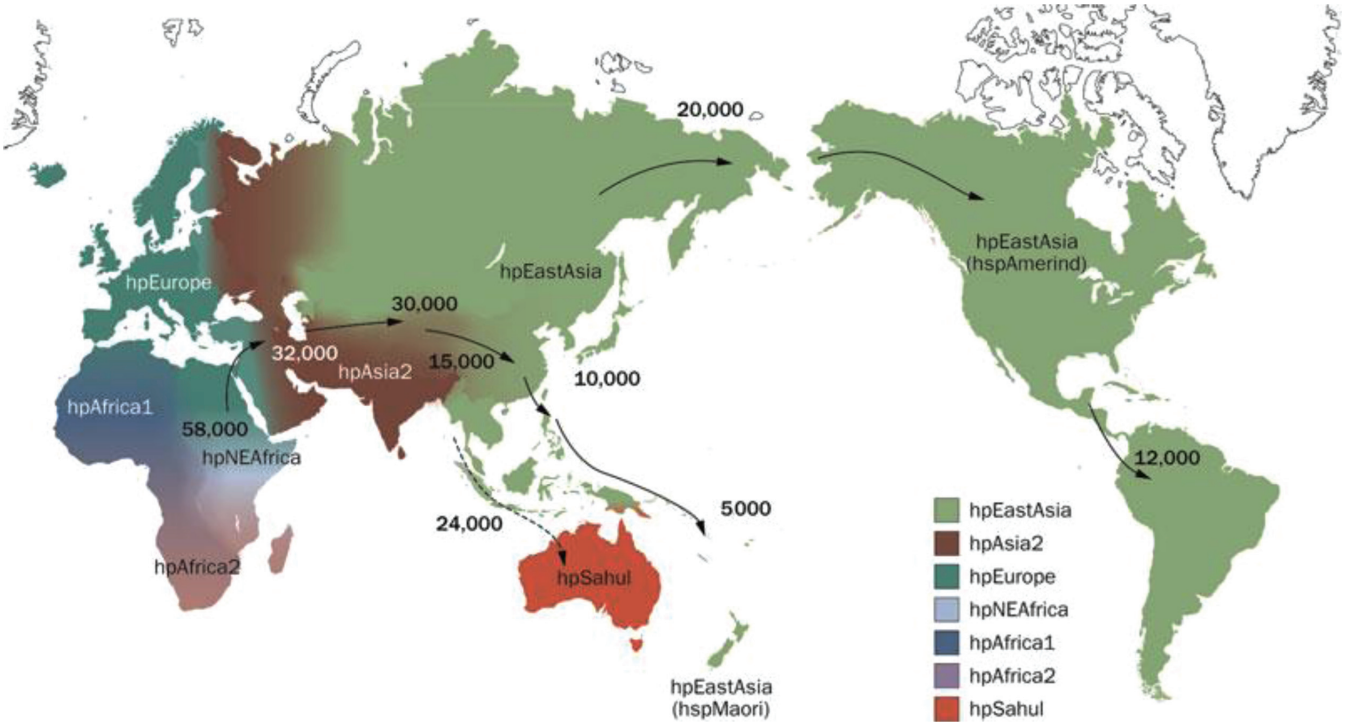


FIGURE 20.3

Origin and spread of *H. pylori* strains around the world. Reprinted with permission from Yamaoka, Y., "Mechanisms of disease: *Helicobacter pylori* virulence factors," *Nature Reviews Gastroenterology and Hepatology* 7:629, 2010.

2005—not a bad reward for making yourself sick by infecting yourself with bacteria!

Getting back to anthropology, studies of genetic variation in *H. pylori* have revealed the existence of several different strains, and interestingly, the spread of these *H. pylori* strains around the world (Figure 20.3) mirrors quite nicely current thinking concerning the spread of modern humans around the world. In particular, *H. pylori* seems to have arisen in Africa around 100 kya and spread from Africa first via an early southern dispersal about 60 kya that gave rise to unique strains in Sahul (the combined Australia–New Guinea landmass), with a separate, later origin for the major strains in Europe and Asia (Falush et al. 2003; Linz et al. 2007). Moreover, New World strains of *H. pylori* are derived from Asia, and the *H. pylori* found in Polynesians seems to have a Taiwanese origin (Falush et al. 2003; Moodley et al. 2009). These findings are all in excellent agreement with the current views on the origins and dispersals of modern humans, as discussed in Chapter 16. So, the genetic variation in *H. pylori* (and, perhaps, other commensal—that is, human associated—bacteria) can be used as an independent source of information about human migrations.

Moreover, there is good reason to think that bacteria may be able to shed light on demographic events that are too recent to leave a signal in human genetic variation. For example, a study of some communities in India found no significant differences among them based on human genetic variation, but there were significant differences based on *H. pylori* variation (Wirth et al. 2004). Bacteria evolve much faster than we do because of their much shorter generation time (typically measured in hours, not years), and so it is not unreasonable to think that genetic differences can accrue among bacteria more quickly than genetic differences among their hosts. Of course, differences in *H. pylori* strains among these communities could also reflect some environmental or dietary difference, but the point is that the history of these communities has not been identical in all aspects, even though we can't distinguish any genetic differences among them.

So, one interesting anthropological use of bacteria is to see what they can tell us about human migrations and population relationships. The other potentially interesting anthropological use of bacteria arises from the growing realization that bacteria do more than just live in more or less harmless association with us (except when they cause disease). Bacteria can play an active role in aspects of our health, diet, and perhaps even our behavior. Regarding health, see Figure 20.4. The two mice in the photograph differ by a genetic mutation that causes the mouse on the right to become obese. But their gut microbiomes also play a role in the obese phenotype: take germfree, nonobese mice

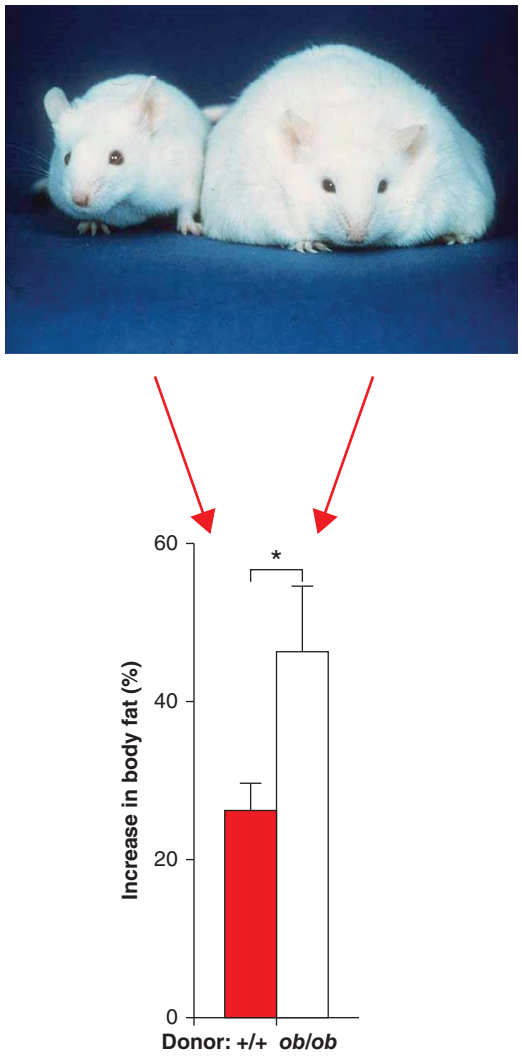


FIGURE 20.4

Top, the photograph shows a mouse homozygous for a mutation that results in obesity, compared to a normal laboratory mouse. Bottom, increase in body fat observed when the gut microbiome from normal (+/+) or obese (ob/ob) mice is introduced into germfree mice. There is a significant increase in body fat associated with the gut microbiome from obese mice. Top, modified with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Fatmouse.jpg); bottom, modified with permission from Turnbaugh, P.J., et al., "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature* 444:1027, 2006.

(i.e., "normal" mice that have been raised in a sterile environment) and infect them with the gut microbiomes of obese and nonobese mice, and the mice infected with the gut microbiomes of the obese mice become fatter than the mice infected with the gut microbiomes of the nonobese mice. Studies have also found characteristic differences in the gut microbiomes of obese versus nonobese humans that tend to disappear when the obese people lose weight (reviewed in Ley 2010). I have to confess that I rather like the implications of these results—if I am overweight, maybe it's not because I eat too much or don't exercise enough, maybe it's because I have fat bacteria!

Regarding diet, it turns out that some of the bacteria living in the gut of many Japanese people have acquired genes that make it easier for them to digest seaweed; bacteria living in the gut of other people generally don't have these genes (Hehemann et al. 2010). The researchers who conducted this study hypothesized that frequent consumption of sushi, which is often wrapped in seaweed, exposed the gut bacteria of Japanese to marine bacteria that could digest seaweed. The gut bacteria were able to incorporate the seaweed-digesting genes from the marine bacteria via horizontal transfer and thereby gain the valuable ability to digest seaweed—at least, valuable to them since their hosts regularly ate seaweed. As one commentator noted, this study gives new meaning to the idea that non-Japanese don't have the guts to eat sushi! Anyway, you can be sure that there will be many more studies examining the influence of variation and evolutionary changes in our diet (such as the development of agriculture) on both us and our microbiomes.

Finally, regarding behavior, there is evidence to suggest that bacteria can manipulate the behavior of their hosts. For example, bacteria can influence the mating behavior of fruit flies. It's well known that if you take some fruit flies and raise them on molasses, take other fruit flies and raise them on starch, and then put them together in mate choice experiments, then the fruit flies will mate preferentially with flies grown on the same food source (Figure 20.5). Why this happens was a mystery—until microbiologist Eugene Rosenberg decided to investigate the microbiome, and sure enough, it turns out that the gut microbiomes are responsible (Sharon et al. 2013). Treat the fruit flies with antibiotics to kill off their gut bacteria and the mating preference disappears; infect fruit flies raised on starch with the microbiome of flies raised on molasses, and now the "starch" flies prefer to mate with "molasses" flies. From the bacteria's point of view, this behavior makes eminent sense—if you're a bacterium adapted to molasses as a food source, then you want your host's offspring (in which your descendants will live) to continue to utilize molasses as a food source and not mess around with flies that eat something else. How the bacteria accomplish this isn't known, although there is some evidence to suggest that the pheromones (hormones involved in sexual attraction) differ between "starch" and "molasses" flies, so the bacteria may be manipulating the expression of pheromones.

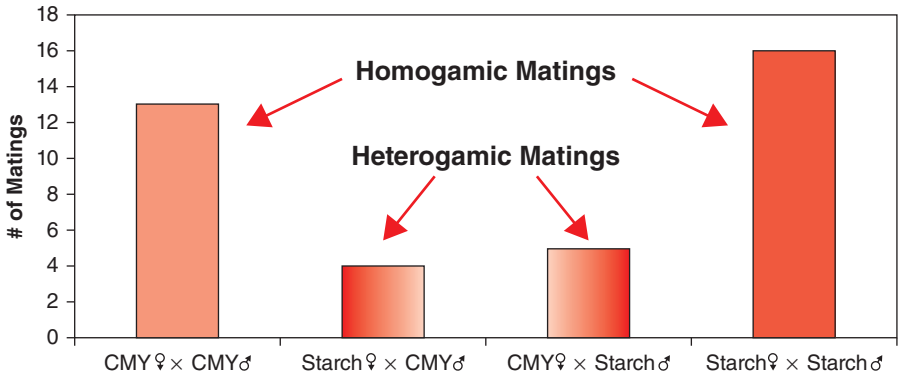


FIGURE 20.5

Relative frequency of matings involving fruit flies raised on either starch or molasses (CMY) as a food source. Homogamic matings involve flies raised on the same food source, while heterogamic matings involve flies raised on different food sources. The figure shows that flies prefer to mate with flies that were raised on the same food source. Modified with permission from Sharon, G., et al., "Commensal bacteria play a role in mating preference of *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences USA* 107:20051, 2010.

You might be thinking that human behavior is a little far-removed from fruit flies, but other evidence of behavioral manipulation comes closer to home. The parasite *Toxoplasmosis gondii* (which is not actually a bacterium but is pretty close as it is a single-celled organism) can infect pretty much any mammal but needs to infect a cat in order to complete its life cycle. In humans, infection results in a disease called toxoplasmosis, which is usually fairly benign, unless your immune system is compromised (e.g., due to AIDS) or if you are pregnant (in which case, there is a high risk of miscarriage, birth defects, or even death of the infant). Infected cats (and only cats) shed parasite eggs in their feces—this is why pregnant women should avoid cleaning a cat's litter box. In the wild, rodents frequently become infected via contact with infected cat feces; infected rodents then develop cysts in their brains but otherwise do not transmit the infection to others (unless they get eaten). Intriguingly, it looks as if infected rodents behave differently than noninfected rodents—and in such a way as to increase the likelihood that they will be eaten by cats! This was shown in a study in which rats that were experimentally infected with the toxoplasmosis parasite were then placed in outdoor enclosures with four different scented areas, consisting of untreated straw, straw scented with the rat's own urine, straw scented with rabbit urine, and straw scented with cat urine (Berdoy et al. 2000). The number of times each rat visited each scented area was then recorded, and the only significant difference between infected and noninfected rats was in the cat-scented area: infected rats visited the cat-scented area nearly twice as often as did noninfected rats (Figure 20.6)! The title of the study reporting these results nicely sums up this finding: "Fatal attraction in rats…". Again, this sort of behavioral modification makes sense if you're a toxoplasmosis parasite: if you need to infect a cat in order to reproduce, then if you find yourself in a different host, anything you do to facilitate a cat eating that host will be selected for. How the parasite manages this behavioral modification is still not known, but the existence of cysts in the brains of infected rodents certainly suggests that they could somehow modify brain gene expression. There is increasing evidence that the gut microbiome can influence brain development and behavior (reviewed in Mayer et al. 2014), leading to all sorts of wonderful benefits to humans (this last sentence was edited by my gut microbiome…).

Overall, these sorts of studies tend to support the view that selection does not just operate on us and our genome, it operates on what has been called the "hologenome"—that is, our genome plus our microbiome (Zilber-Rosenberg and Rosenberg 2008). If some component of our microbiome can act on us in such a way as to gain an advantage in terms of survival and/or reproduction, then natural selection will favor whatever it is causing us to do. Perhaps in the future we'll see defense attorneys claiming that an individual's microbiome was responsible for his or her criminal actions, or students claiming that their microbiome is to blame for them not doing their homework! At any rate, investigations of the microbiome, the role it plays in health and disease, and the anthropological implications and applications are all very active areas of research that promise to be very fruitful indeed.

MORE ANALYSES

The deluge of data that we are already experiencing, with the promise of more to come, is of little use if

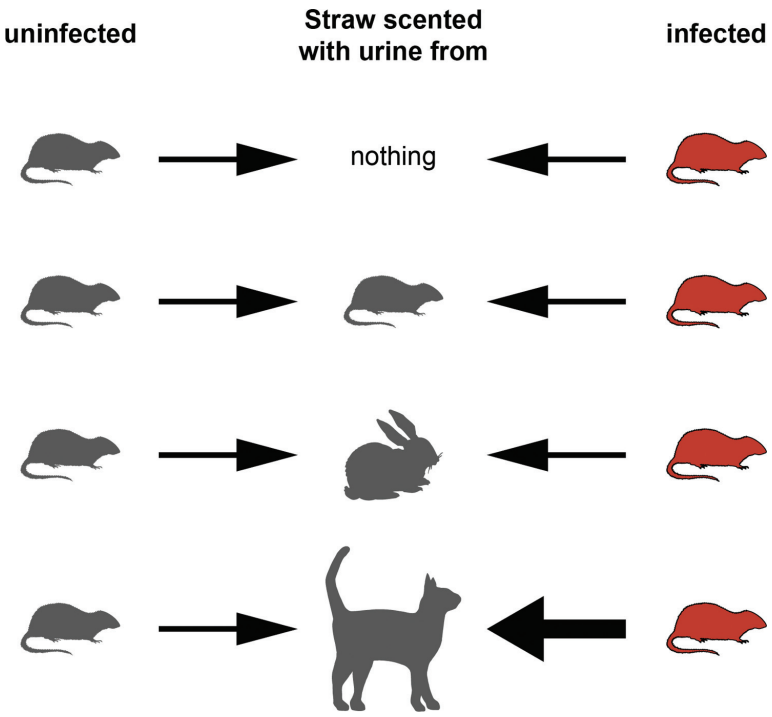


FIGURE 20.6

Relative amount of time rats that are either uninfected or infected with *Toxoplasma gondii* spend investigating straw that is scented with either nothing (unscented), rat urine, rabbit urine, or cat urine. The only significant difference between infected and uninfected rats is that infected rats spend significantly more time investigating straw scented with cat urine (as indicated by the larger arrow for the infected rats). Data from Berdoy, M., et al., "Fatal attraction in rats infected with *Toxoplasma gondii*," *Proceedings of the Royal Society B* 267:1591, 2000.

we don't know what to make of it all. Fortunately, as has turned out to be the case in the past, when new technical developments make new kinds of genetic (or nowadays, genomic) data available, the population geneticists and computational biologists quickly start investigating how they can make use of such data. Methods for visualizing overall patterns in genome-wide data (e.g., PCA and STRUCTURE-like analyses, discussed in Chapter 11) are now routinely employed, and a current focus of much ongoing work is how to make inferences about demographic history from such data—that is, the history of population size changes, population splits, migration and admixture, and so forth. We've touched upon some such methods that use simulations or other ways to infer historical events and estimate parameters of interest, but you can be sure there is lot more to come. The PSMC (pairwise sequential Markovian coalescent) method for inferring the history of population size change from a single genomic sequence (mentioned in Chapter 12) is just one example of the newest approaches, and already a new version for comparing multiple genome sequences (called, imaginatively, MSMC for multiple sequential Markovian coalescent) has been developed (Schiffels and Durbin 2014), with improved performance for inferring population size changes in recent times. Other approaches currently under investigation make use of patterns of linkage disequilibrium (LD)—associations among linked polymorphic sites—as there is much additional information about demographic history in these associations that is not captured by the "traditional" approaches based on the allele frequency spectrum and the like. That is, you gain more from the associations than you do by simply analyzing your data as a lot of single polymorphisms—in other words, the whole is greater than the sum of the parts.

One area in which much is being done, but much more needs to be done, is in detecting, quantifying, and dating admixture events. As we have seen, one of the major contributions of the molecular approach to anthropology is the demonstration that archaic humans have contributed genetic ancestry to modern humans. Getting more precise estimates as to how much Neandertal/Denisova admixture there is in our genomes (and how much this varies among individuals), dating the admixture events, and figuring out if the archaic admixture contributed genetic variation of adaptive value to our ancestors—these are all important and interesting areas of current intensive investigation. And, there is of course the question as to what other archaic human groups our ancestors might have interbred with. Indeed, some recent studies have claimed to find evidence of interbreeding between some as yet unknown archaic human group and Africans (Hsieh et al. 2016). However, as discussed in Chapter 14, without an actual genome sequence from the archaic group in question, it is extraordinarily difficult to distinguish between archaic interbreeding and other explanations (such as ancient population structure) for "weird" signals in our genomes—you should keep in mind that most of the previous claims for genes showing Neandertal ancestry in modern humans were not substantiated when the Neandertal genome sequence became available.

And admixture is not just important in the context of archaic human ancestry—as we get better and better at detecting admixture between different human populations in genome-wide data, we are finding more and more evidence of it. Simple admixture models are fine when it comes to something like detecting the African versus European ancestry in African–Americans, for example, but often the situation is not so simple—multiple admixture events are probably the rule rather than the exception. Many Latin-American groups have ancestry from native Americans, Europeans, and Africans (Moreno-Estrada et al. 2013); a recent study of Indonesian populations (Lipson et al. 2014) identified four distinct ancestries in some populations (corresponding to probable Papuan, Filipino "Negrito," mainland southeast Asian, and Taiwanese origins); and recent work on Siberian populations has identified up to six distinct ancestries in some groups (Pugach et al. 2016). So, detecting, quantifying, and dating the various admixture events that have occurred among different human populations are also an area of intense interest and investigation, and new methods to do so are coming out all the time.

Still, an important limitation of current methods for analyzing admixture is that you have to have reasonable guesstimates for the ancestral, parental populations that participated in the admixture. That is, you don't always have to have the direct descendants of the parental groups involved in the admixture, but descendants of a group that is fairly closely related to the admixing group are still needed to be able to analyze the admixture signal. But sometimes we might not have this. For example, molecular anthropologist and linguist Brigitte Pakendorf and colleagues (including myself) recently investigated an mtDNA haplogroup that is characteristic of southern African Khoisan foragers (i.e., hunter-gatherer groups who speak languages with click consonants) and found distinct, divergent lineages of this haplogroup that were restricted to Bantu-speaking groups from Zambia (Figure 20.7). These divergent lineages are not found in Bantu-speaking groups elsewhere, nor are they found in any extant Khoisan group (which we can say with some confidence, since essentially all existing Khoisan groups have been analyzed). The most likely explanation for their presence in Zambian Bantu-speaking

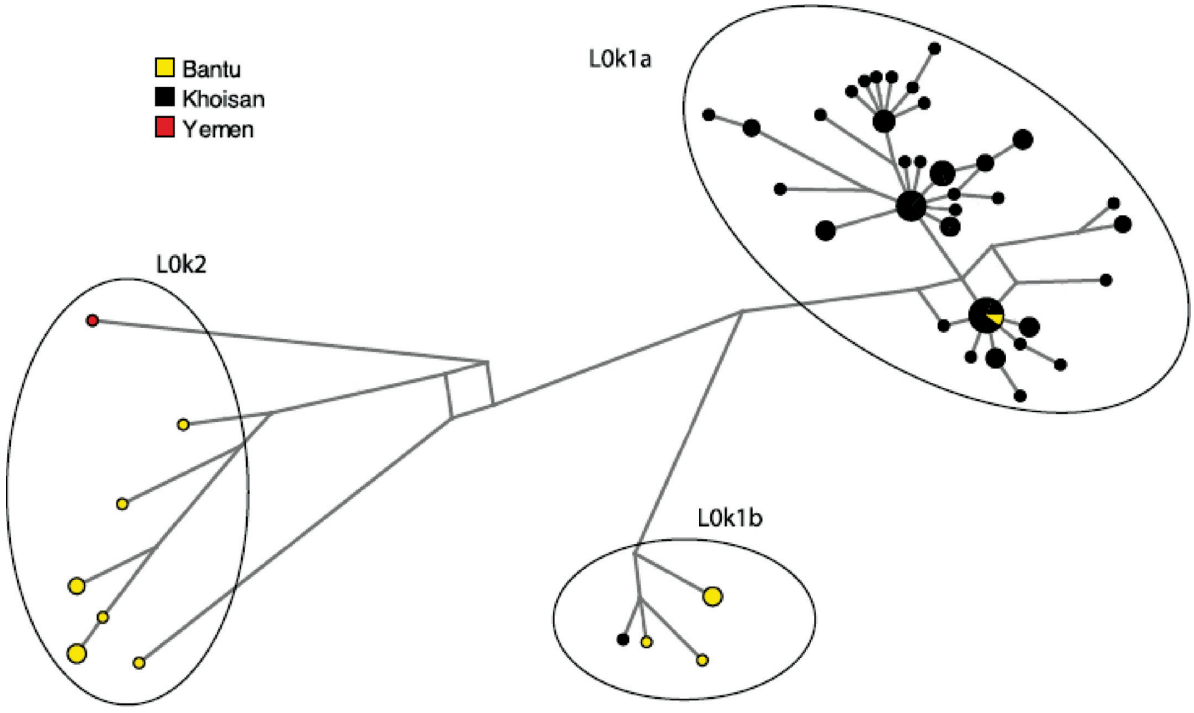


FIGURE 20.7

Network of mtDNA genome sequences belonging to haplogroup L0k in southern African populations. L0k is considered to be of Khoisan origin, and indeed the most frequent and most diverse lineage, L0k1a, is almost exclusively found in Khoisan groups. However, lineages L0k1b and L0k2 are almost exclusively found in Bantu-speaking individuals. Moreover, these Bantu groups are from areas of Zambia where no Khoisan groups exist today. Reprinted with permission from Barbieri. C., et al., "Ancient substructure in early mtDNA lineages of southern Africa," *American Journal of Human Genetics* 92:285, 2013.

groups is that when Bantu speakers arrived in Zambia some 2200 years ago, they admixed with a resident Khoisan group that subsequently went extinct. Linguistic evidence supports this scenario, in that some of these Zambian Bantu languages have a few words with click consonants that were undoubtedly borrowed from a Khoisan language—and while some of these can be traced to existing Khoisan languages, others cannot be derived from any known Khoisan language. So, not only do we carry the genetic ancestry of extinct archaic humans in our genome, we can also find genetic traces of extinct modern human groups in current human populations.

There are probably many other human groups that went extinct, especially as a consequence of agricultural expansions (such as the Bantu expansion), but we can potentially recover some information about their genetic relationships and thus increase our knowledge about our prehistoric genetic structure. Indeed, we already know of some such groups; for example, there are no longer any "full-blooded" native Tasmanians, but there are people with Tasmanian ancestry, so the potential exists to learn about the genetic ancestry of Tasmanians. However, we first need to come up with the methods to allow us to detect and analyze the genetic ancestry signal of extinct groups, especially in the absence of suitable proxies among current groups, and this remains a very difficult problem. Still, some clever people are hard at work on this, so hopefully some workable solution(s) will be forthcoming.

In addition to aspects of demographic history such as population size changes, migration/admixture, and so forth, there is also room for improvement in methods for detecting selection, both species-wide and local selection. In general, with current methods we have reasonable power to detect strong, repeated (i.e., selection for multiple mutational events), and/or recent selection, but anything more subtle won't be picked up. From one perspective, this is not such a big drawback, because as we saw back in Chapter 18, we still have a hard time with trying to decipher those signals of selection that we do detect. In particular, given the typical result from a genome scan for signals of selection—namely, a laundry list of potential candidate genes for selection—it remains a difficult task to sort out the real signals from the false positives and then figure out what is behind the real signals.

But from another perspective, this is an important issue: there is growing suspicion that the classic model of a selective sweep presented previously (and repeated in the left side shown in Figure 20.8) may happen only rarely. That is, in this model, there is some selective force (e.g., exposure to lactose in adulthood from drinking cow's milk), and then a new mutation arises that happens to confer an advantage with respect

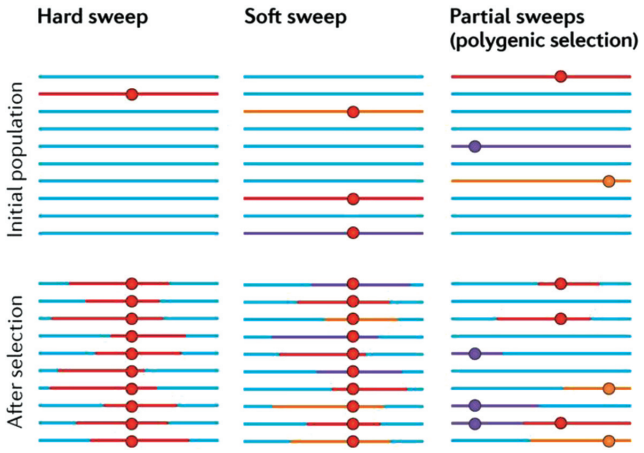


FIGURE 20.8

Difference between hard sweeps, soft sweeps, and selection involving polygenic adaptation. Left, a new variant (red circle) arises and sweeps to fixation, along with a long associated haplotype (red bars). Middle, a preexisting variant (red circle) on a variety of haplotypes (colored lines) is selected for, so there is no association of the variant with a specific haplotype. Right, different variants are selected on different haplotypes, with the result that any signal of selection is dispersed across the genome. Reprinted with permission from Cutter, A.D., and Payseur, B.A., "Genomic signatures of selection at linked sites: unifying the disparity among species," *Nature Reviews Genetics* 14:262, 2013.

to this selective force (e.g., persistence of lactase into adulthood), so then selection causes the new mutation to increase rapidly in frequency, producing the characteristic signature of a selective sweep. But according to an alternative model (middle of Figure 20.8), advantageous mutations may already exist in the population as neutral polymorphisms, before circumstances change to make them advantageous. Before selection happens, such mutations are subject to the whims of genetic drift, slowly increasing and decreasing in frequency, recombining onto new haplotypes, and so forth. Then, when selection starts to happen, there may be several haplotypes that increase in frequency, rather than a single haplotype as assumed under the selective sweep model. Selection on preexisting mutations (also known as **standing variation**) leads to what is called a **soft sweep**, and as you might imagine from looking at Figure 20.8, it is very difficult to distinguish the signal of a soft sweep from neutral expectations. Moreover, it may be that soft sweeps happen much more frequently than hard sweeps—after all, if circumstances change and there is a new selective force, then any preexisting mutation that happens to now be advantageous will automatically be selected for, which

might then make it more difficult for subsequent, new advantageous mutations to ever get started and increase in frequency. And, as you can also see from Figure 20.8, things get even more difficult if multiple genes are involved in a particular adaption (which is called **polygenic adaptation**). So, even though we may be tempted to focus on hard sweeps just because they are easier to detect, it could very well be that in doing so, we are excluding from consideration most of the selective events that have actually happened. Anyway, developing effective methods for detecting soft sweeps and polygenic adaptation—as well as further refining methods for detecting hard sweeps—is an active area of research that will hopefully bring new insights into the role of selection during human evolution.

RELATING PHENOTYPES TO GENOTYPES

Back before genetic analyses became possible, anthropological studies of human population variation had to rely on physical characteristics—things like skin pigmentation, eye and hair color, body size and stature, limb proportions, dry versus wet ear wax, and so forth. Indeed, this is where the term "physical anthropology" comes from. These studies not only documented the variation in physical characteristics in various populations but also sought general principles to help explain this variation. For example, the variation in body size and limb proportions among human populations generally conforms to what are known as Bergmann's and Allen's Rules. Bergmann's Rule, named after German biologist Carl Bergmann, holds that individuals with higher body mass tend to be found in colder climates (and thus further from the equator), while individuals with lower body mass tend to be found in warmer climates (and thus closer to the equator). Allen's Rule, named after American zoologist/ornithologist Joel Allen, is based on the observation that individuals from warmer, equatorial climates tend to have longer limbs than individuals from colder, polar climates. Both of these rules have to do with conservation of body heat in colder climates and loss of body heat in warmer climates. The idea is that there is an optimal amount of body heat for people to maintain—too little, and you succumb to hypothermia, too much, and you suffer from heat exhaustion or heat stroke. And retention/loss of body heat is related to the amount of exposed surface area. So, a compact body shape with short limbs is optimal for retention of body heat (think of Eskimos, for example) as surface area is minimized, while an elongated body shape with long limbs is optimal for shedding body heat (think of Dinkas or Maasai from East Africa, for example) because surface area is maximized (Figure 20.9).

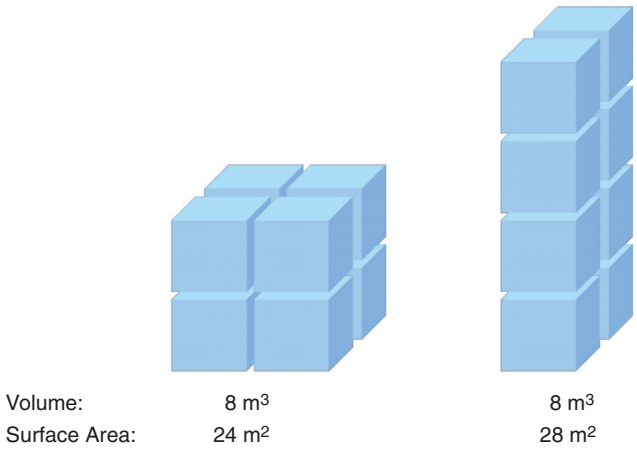


FIGURE 20.9

Compact shape has a smaller surface area than a more elongated shape with the same volume. Thus, compact body shapes will conserve body heat and are selectively advantageous in cold temperatures, while elongated body shapes will lose body heat more quickly and hence are selectively advantageous in warm temperatures. Modified with permission from Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Volume_surface.svg).

In humans, body mass is indeed negatively correlated with mean annual temperature, in accordance with Bergmann's and Allen's Rules, although the strength of the correlation has weakened considerably during the past 50 years or so (Katzmarzyk and Leonard 1998). This decrease in correlation seems mostly due to increased body mass in individuals living in warmer climates, which probably reflects changes in diet associated with "Westernization."

You might think that such differences in body proportions surely reflect genetic differences between populations. After all, if a Dinka couple moves to Alaska, they are not going to start having children with the body shape of Eskimos. But there is good evidence that some body shape variation among humans is indeed influenced by the environment. Native Americans who live at high altitude in the Andes frequently develop a characteristic "barrel-shaped" chest, and it turns out that this is not due to genetics: if barrel-chested individuals move to low altitude and have children, their children will develop normal chests (Frisancho and Baker 1970). So, this is not a genetic trait, but rather a physiological change that occurs in some individuals raised from childhood at high altitude, presumably in response to the lower amount of oxygen at high altitude.

This is an example of the major issue that arises with studies of phenotypic variation: which of the various possible influences on such variation (genetic,

climate, diet, etc.) are most important in explaining the differences among populations? Moreover, when we observe the same phenotypic trait in different populations, is the genetic basis also the same or is it different (think, for example, of lactase persistence in European vs. African populations—same phenotype, but different mutations). Clearly, understanding the genetic basis of phenotypic variation would aid substantially in interpreting such variation. And there has been considerable progress in understanding some of the genes involved in some phenotypic traits, particularly skin, hair, and eye pigmentation, so let's take a look at what we've learned about these "colorful" traits.

Why there should be such striking variation in human skin pigmentation around the world has attracted attention at least since Aristotle's time, and many explanations have been proposed. Darwin favored sexual selection, but while one should always keep in mind the potential for sexual selection when dealing with any physical attribute of humans, most researchers do think that skin pigmentation variation is the result of natural selection. In particular, it's long been known that skin pigmentation variation is highly correlated with levels of exposure to ultraviolet (UV) light, which in turn tend to be higher in equatorial (and high altitude) regions of the globe (Figure 20.10). While in principle this correlation could reflect association rather than causation, current thinking is that variation in UV exposure has directly influenced variation in human skin pigmentation. There are thus two aspects to consider: why is dark skin pigmentation favored in areas of high UV exposure, and why is lighter skin pigmentation favored in areas of low UV exposure?

With regard to the first, dark skin pigmentation probably first evolved in our ancestors shortly after they lost body hair. We can infer this because all other apes have light skin pigmentation underneath their hair and develop darker pigmentation only on body areas that lack hair. As discussed in Chapter 19, loss of body hair in our ancestors probably facilitated thermoregulation (in particular, loss of body heat) and occurred early in our evolution, soon after our ancestors moved into a more open, Savannah-like environment and became more mobile. This loss of body hair led to changes in the structure of the skin that probably increased resistance against parasites that found hairless skin particularly appetizing. As to why darker skin pigmentation also evolved after the loss of body hair, several explanations have been put forward (Jablonski 2006). For example, darker skin pigmentation may also play a role in parasite resistance, as **melanin** (the main pigment present in skin) has been shown to inhibit bacterial/fungal infection and growth (Mackintosh 2001). It is also quite clear that people with darker skin have a much lower risk of developing skin cancer from exposure to sunlight, whereas people with lighter skin have a much higher risk—as those who have moved from Europe to sunnier climes, like Australia or Africa, have learned the hard way. However, it does not seem as if decreased susceptibility to skin cancer alone can explain selection for darker skin in high UV areas. Most skin cancers are not fatal or even terribly debilitating, and even those that do have severe consequences occur too rarely among people of reproductive age to have any significant impact on reproductive success (although one should of course be cautious in extrapolating from what happens when contemporary Europeans head to sunnier climes—circumstances for our ancestors on the African Savannah a few million years ago may have been quite different).

The most plausible selective advantage for dark skin pigmentation seems to be related to the degradation of vitamin B9, also known as folic acid or folate. Folate is an essential vitamin (meaning that our bodies cannot make it, so we have to get it from our diet) that plays an important role in DNA synthesis and repair, and folate deficiency is associated with an increased risk of neural tube defects, or NTDs for short (reviewed in Borradale and Kimlin 2012). Neural tube defects are among the most common birth defects, with an occurrence of about 1 in 1000 births, and involve abnormalities in the formation and closure of the neural tube (brain and spinal cord). There are a variety of NTDs that range in severity, but in many cases the affected infant dies soon after birth or is paralyzed. Supplementing the diet of pregnant women with folate led to a significant decrease in births with NTDs, and many countries now routinely fortify flour with folate as a convenient way of ensuring adequate amounts of folate in the diet. And more recently, it has been discovered that exposure to UV light can increase folate degradation (reviewed in Borradale and Kimlin 2012), which therefore could lead to a higher risk for NTDs. So, according to this hypothesis, darker skin pigmentation would have been selected for after we lost body hair as a means of decreasing degradation of folate and hence reducing the risk of NTDs. This hypothesis is very attractive, because it stands to reason that anything that decreases the risk of severe birth defects will be selectively advantageous. Still, while there is some preliminary evidence that dark skin pigmentation does reduce folate degradation, and moreover that people with dark skin pigmentation have an overall lower risk of NTDs, there is still much more that needs to be done to investigate and verify the folate degradation hypothesis as the explanation for the selective advantage of dark skin pigmentation.

And what about lightening of skin pigmentation away from the equator? Here we seem to be on somewhat firmer ground, as there is a fairly well-established

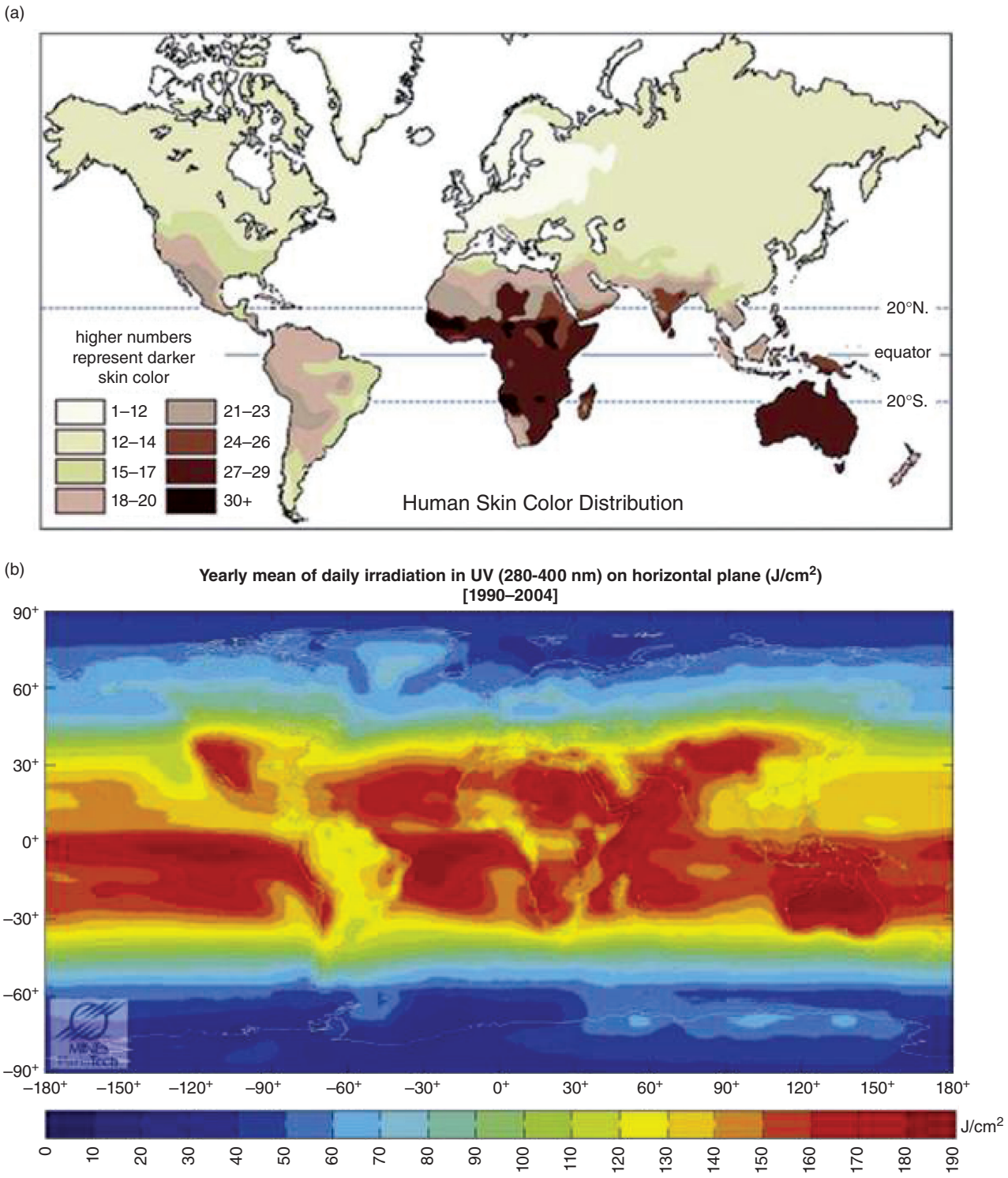


FIGURE 20.10

Variation in skin pigmentation (a) and UV dose (b) around the world. UV, ultraviolet. Reprinted with permission from Liu, F., et al., "Colorful DNA polymorphisms in humans," *Seminars in Cell & Developmental Biology* 24:562, 2013.

hypothesis centered on another vitamin, namely, vitamin D. Vitamin D deficiency leads to rickets, a disease that mostly affects children and involves soft bones that are prone to growth deformities and fractures. Vitamin D is, strictly speaking, not an "essential" vitamin because most mammals—including humans—are able to make vitamin D upon exposure to sunlight. In the 1920s, the biochemist Harry Steenbock established a rat colony in the Department of Agricultural Chemistry at the University of Wisconsin—over the objections of colleagues who spent a lot of time advising farmers how to get rid of rats—and showed that rickets in rats could be prevented not only by exposing the rats to UV but also by exposing their food to UV (Steenbock 1924). This UV exposure was sufficient to convert the precursor of vitamin D in the rat food to vitamin D, which suggested to Steenbock that rickets in humans could similarly be prevented by irradiating food for human consumption. Unable to convince the university administrators of the potential commercial importance of this discovery, Steenbock spent $300 of his own money to take out a patent. His foresight was rewarded when the Quaker Oats company then offered him a million dollars for the patent rights. However, Steenbock was an idealist who believed that the university should manage and profit from his discovery, not companies. So, with some fellow alumni he set up the first university technology transfer office (viz., the Wisconsin Alumni Research Foundation, still in operation today), which then licensed the technology to several companies. The result was that it became common practice to irradiate various foods (most notably milk, as it was illegal at the time to add any substances to milk, but not to irradiate it), thereby ensuring that everyone received adequate amounts of vitamin D, and rickets was largely eliminated in the United States by the time the patent expired in 1945.

What does vitamin D have to do with light skin pigmentation? It turns out that dark skin pigmentation inhibits vitamin D synthesis. In sunny climes this is not an issue, because there is ample sunlight and UV exposure, enough to ensure that adequate amounts of vitamin D are synthesized even by people with very dark skin pigmentation. But in areas that receive less sunshine, such as Europe or northern Asia, people with dark skin pigmentation do tend to have more of a problem with vitamin D deficiency and the associated consequences (like rickets and a higher risk of miscarriages) than people with light skin pigmentation. So, the vitamin D hypothesis holds that after modern humans left Africa for more northerly climes, they would have experienced selection for lighter skin pigmentation in order to facilitate synthesis of adequate amounts of vitamin D.

While this is the generally accepted view, it is important to keep in mind that the actual data in support of this hypothesis are not as solid as one might expect for this "textbook" explanation. Initial studies that suggested that people with dark skin pigmentation living in northern climates suffered more from rickets than people with light skin pigmentation did not take into account the fact that people with dark skin pigmentation in northern climates tended to live in much worse circumstances (inner-city slums or the like) than people with light skin pigmentation; when differences in living conditions were controlled for, differences in the prevalence of rickets in people with dark versus light skin pigmentation largely disappeared (Robins 2009). However, overall there does seem to be some influence of skin pigmentation on vitamin D insufficiency (Chaplin and Jablonski 2009), even if we don't fully understand all of the health consequences of vitamin D insufficiency.

So, to summarize, the association between levels of UV exposure and skin pigmentation (Figure 20.10) is potentially explained by selection for dark skin pigmentation in areas with high UV exposure to prevent degradation of folate and selection for light skin pigmentation in areas with low UV exposure to facilitate vitamin D synthesis. What does genetics have to add to this story? Figure 20.11 outlines some of the major steps involved in the synthesis of the two major types of melanin (skin pigment) in humans, namely, **eumelanin** and **pheomelanin**, and the key genes involved. The *MC1R* gene, which regulates an

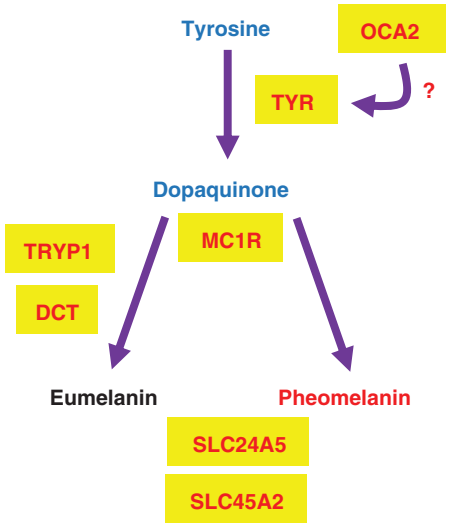


FIGURE 20.11

Some of the important steps and genes (yellow boxes) involved in the synthesis of the two types of melanin in humans, eumelanin (brown/black) and pheomelanin (red). The *MC1R* gene product is thought to be involved in both pathways, while the *SLC24A5* and *SLC45A2* gene products are probably involved in transport of these skin pigments.

important step in determining the production of eumelanin and pheomelanin, was one of the first such genes to be analyzed. Compared to chimpanzee MC1R, there does appear to be an excess of amino acid substitutions during the evolution of human MC1R, although it is not clear whether this excess is really a significant signal of selection on human MC1R (Rogers et al. 2004). Interestingly, this gene shows practically no nonsynonymous polymorphisms in Africans, but several nonsynonymous polymorphisms are known in non-Africans, and some of these nonsynonymous polymorphisms are associated with red hair color (Harding et al. 2000). It thus appears that MC1R may have been subject to selection in humans after our lineage diverged from that of apes, possibly for dark skin pigmentation after our ancestors lost body hair (Rogers et al. 2004). Moreover, MC1R appears to be under strong functional constraints in Africans, meaning that essentially no amino acid changes that interfere with the function of the protein can be tolerated, which fits nicely with the idea that there has been selection for dark skin pigmentation in Africa. Furthermore, the elevated level of nonsynonymous polymorphisms outside of Africa suggests reduced functional constraints on MC1R, meaning that reduced MC1R function can be tolerated outside of Africa. Reduced functional constraints on MC1R outside of Africa fits nicely with the idea that there is no need to maintain dark skin pigmentation outside of Africa; however, there is no signal of actual selection on MC1R to produce the lighter skin pigmentation in Europe and Asia.

Instead, other pigmentation-related genes show signatures of selection outside of Africa. There is a strong signature of selection in both Europeans and Asians in a large genomic region that includes a gene called *KITLG* (Williamson et al. 2007), which is involved in regulating the production and maintenance of **melanocytes** (cells that synthesize melanin). This selective sweep signal seems to be driven not by polymorphisms in the *KITLG* gene itself but rather by SNPs that are upstream from the gene and may influence regulation of KITLG expression. For example, 326kb from the *KITLG* gene there is a SNP (called rs642742—don't ask why) in which the derived allele is present at frequencies of greater than 80% in Europeans and Asians but at less than 10% in Africans. That this SNP could have something to do with skin pigmentation variation was deduced not from the selection signal but rather from a study that found that pigmentation variation in stickleback fish is influenced by polymorphisms that in turn influence the regulation of KITLG expression in these fish (Miller et al. 2007). This study also showed that in African–Americans the rs642742 SNP shows a significant association with skin pigmentation: individuals homozygous for the derived allele (at high frequency in Europeans and Asians) have paler skin color on average than individuals homozygous for the ancestral allele (at high frequency in Africa); heterozygotes tend to have intermediate skin colors. Although the rs642742 SNP is located in a noncoding region that is highly conserved in mammals, suggesting that it could have a direct effect on KITLG expression, it is also possible that the rs642742 SNP results reflect association rather than causation—that is, the SNP could be in strong LD with some other (unknown) genetic variant that is truly responsible for the skin pigmentation variation.

KITLG seems to be the exception rather than the rule among non-African populations when it comes to selection for lighter skin pigmentation, in that it is the only pigmentation variation gene found so far in which the signal of selection is shared by European and Asian populations. Selection signals at all other skin pigmentation variation genes are largely specific to either European or Asian populations. For example, the genes *SLC24A5* (which was first identified as a contributor to skin pigment variation in humans after a mutation in this gene was found to underlie pigment variation in zebrafish (Lamason et al. 2005), similar to the KITLG story—there does seem to be something fishy about skin pigmentation studies!), *SLC45A2* (not to be confused with *SLC24A5*!), and *TRYP1* all harbor derived alleles at high frequency only in European populations, show signatures of selective sweeps only in European populations, and have been associated with differences in skin pigmentation between European and non-European populations (reviewed in Sturm and Duffy 2012). Similarly, the gene *OCA2* harbors a derived allele at high frequency only in East Asian populations, shows a selective sweep signature only in East Asian populations, and is associated with differences in skin pigmentation between East Asian and other populations (reviewed in Sturm and Duffy 2012). Another gene involved in skin pigmentation, *DCT*, shows a strong signature of selection in East Asian populations and is thus a candidate gene for skin pigmentation differences between East Asians and other populations (Myles et al. 2007), but so far studies have not been carried out to see whether the mutations present at high frequency in East Asians actually influence skin pigmentation variation (although similar *DCT* mutations in mice do result in lighter coat color).

So, the overall picture that is emerging with respect to skin pigmentation variation is that around the time our ancestors lost body hair (probably a few million years ago), we developed darker skin pigmentation, mediated at least in part by mutations in the *MC1R* gene. The driving force behind this pigmentation change was likely to protect against folate degradation or other effects of high UV exposure. Once modern humans left Africa, there was then selection for lighter skin pigmentation, possibly to enhance production of

vitamin D in response to lower UV exposure. While the initial selection for lighter skin pigmentation happened in a common ancestor of Europeans and Asians, as evidenced by the *KITLG* gene, most of the lightening of skin pigmentation occurred independently and convergently in Europe (involving genes such as *SLC24A5*, *SLC45A2*, and *TRYP1*) and in East Asia (involving genes such as *OCA2* and perhaps *DCT*).

Recently, this scenario of largely independent and convergent selection for lighter skin pigmentation in Europe and East Asia has received further support and refinement from attempts to date when the selective sweeps occurred (Beleza et al. 2013). The selection on the *KITLG* gene is estimated to have occurred about 30,000 years ago, which is within the time frame before the divergence of Europeans and East Asians is estimated to have occurred (somewhere between 25,000 and 40,000 years ago). However, the selection on the *SLC24A5*, *SLC45A2*, and *TRYP1* genes (all involved in lightening of skin pigmentation in Europe) is each dated to 10,000–19,000 years ago, which is well after modern humans arrived in Europe. These more recent dates for the selection in Europe support a refinement of the vitamin D hypothesis for skin pigmentation lightening, which holds that the diet and lifestyle of early modern hunter-gatherers in Europe (along with the lighter skin pigmentation conferred by the mutations in the *KITLG* gene) would have ensured a sufficient supply of vitamin D. It was only with the onset of agriculture, combined possibly with increased use of clothing and shelter as human populations consequently grew in size, that vitamin D insufficiency became a big enough problem that selection for lighter skin pigmentation increased substantially in strength. This is certainly an intriguing hypothesis and may explain why some native Siberian and Alaskan populations have darker skin pigmentation than would be predicted by the UV exposure hypothesis (i.e., because their primarily meat/fish-based diet is richer in vitamin D). However, more research is needed before this hypothesis can be accepted, in particular, dating the selective sweeps on the *OCA2* and *DCT* genes in East Asians.

And ancient DNA studies are providing further insights into the timing and origin of selection events involving skin pigmentation alleles—for example, the allele at *SLC24A5* that is associated with light skin pigmentation in Europeans is absent from early hunter-gatherers in Europe but fixed in early Neolithic farmers in Anatolia, suggesting that the Neolithic migration of farmers from Anatolia to Europe brought this allele to high frequency in Europeans today (Mathieson et al. 2015). Speaking of ancient DNA, what about the skin pigmentation of Neandertals and Denisovans? If lighter skin pigmentation really is advantageous outside Africa, purely because of low UV exposure leading to vitamin D insufficiency in individuals with dark skin pigmentation, then presumably Neandertals and Denisovans would also have experienced selection for this trait. But if instead changes in diet and/or living conditions associated with the onset of agriculture (in combination with low UV exposure) led to vitamin D insufficiency and selection for lighter skin pigmentation outside Africa, then Neandertals and Denisovans should have had dark skin pigmentation. So, knowing whether Neandertals and Denisovans had light or dark skin pigmentation would provide further insights into the reason for the selection for lighter skin pigmentation outside Africa.

This question has been addressed by analyzing the Neandertal and Denisovan genome sequences for all of the mutations known to be associated with lighter skin pigmentation in modern non-Africans. None of these mutations have been found, which has led to the inference that Neandertals and Denisovans had dark skin pigmentation (Cerqueira et al. 2012). However, this is probably not the appropriate analysis to address this question. Given all of the evidence we've already seen for the ubiquity of convergent evolution (different mutations giving rise to the same phenotype) in skin pigmentation variation as well as other traits (such as lactase persistence), we should expect that if Neandertals and Denisovans did have lighter skin pigmentation, it would be the result of novel mutations rather than the same mutations that conferred lighter skin pigmentation in modern humans. Indeed, there is circumstantial evidence in support of lighter skin pigmentation in Neandertals from studies of the *MC1R* gene. Nonsynonymous variants have been found in the *MC1R* gene in Neandertals that are different from those associated with red hair color in modern Europeans but have the same effect on *MC1R* function and so probably also result in red hair (Lalueza-Fox et al. 2007). So, some Neandertals probably also had red hair, and you don't expect such variation in *MC1R* unless they had lighter skin pigmentation as well. Therefore, what is needed to fully address this question of lighter skin pigmentation in Neandertals and Denisovans is to systematically survey all genes involved in skin pigmentation for potential novel variants that would result in lighter skin pigmentation. This task is complicated by all of the previously discussed issues concerning how to associate a mutation with a particular phenotype, especially since in this case we don't even have phenotypic data—we'd have to infer lighter skin pigmentation from some sort of functional assay. Still, with the high-quality genome sequences now available from a Neandertal and a Denisovan, such studies are in progress.

Turning now to variation in eye and hair color, such variation is mostly (but not exclusively) limited to European populations (or populations with European

ancestry). Most people around the world have brown eyes and dark (brown/black) hair, but some have blue or green eyes, and some have light brown, blond, or red hair. Mutations in several genes have been found that are responsible for eye/hair color variation, including some that we've already come across, such as *MCR1* (several mutations are known that cause red or blond hair), *OCA2* (mutations in a gene called *HERC2*, which regulates *OCA2* expression, cause blue/green eye color), and *TRYP1* (a mutation in this gene is associated with blond hair in Melanesians). Selection does not seem to have played a role on these traits, although this could simply reflect the fact that with our current methods we can detect only very strong selection; a weak signal of selection would not be detected.

One important use of this information has been in developing predictive tests for forensic or other purposes. The idea is that with just a DNA sample, how much can you tell about the skin/hair/eye color of the person who left that sample? As it turns out, in some cases quite a bit (Figure 20.12), which is why forensic scientists are keen on seeing how far this can be taken—the ultimate goal would be to be able to produce a complete portrait of a perpetrator or victim just from a DNA sample from a crime scene. That's still a long way in the future, though, for reasons discussed later.

This approach could also be used to predict the phenotype from ancient DNA samples and thereby gain some insights into phenotypic variation in the past. One important caveat, however, is that ancestry also plays a role. For example, one of the first mutations found to be associated with blue eye color in Europeans also occurs in some central Asian populations, where it has nothing to do with eye color variation because there is no eye color variation in these populations—everyone has brown eyes. The reason for this apparent discrepancy is because the mutation does not actually cause blue eye color; instead, it is in strong LD with the true causal mutation. So, what happened is that the associated mutation arose first, spread from Europe to central Asia (or vice versa), and then sometime later the true causal mutation arose in Europe on a haplotype that also carried the associated mutation (Figure 20.13). Thus, in Europe the associated mutation is highly predictive of blue eye color, but in central Asia it tells you nothing about eye color. Clearly, before you could conclude anything about the eye color of an

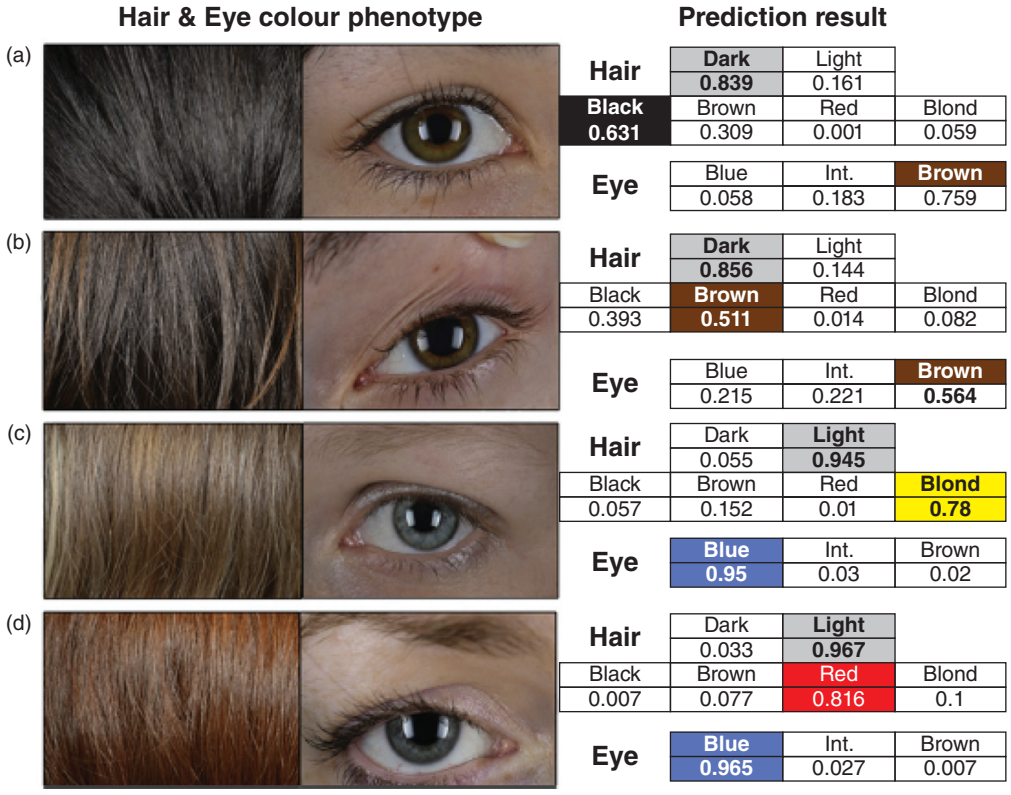


FIGURE 20.12

Example of using targeted genotyping of 24 specific SNPs that either cause or are associated with hair and eye color variation to infer hair/eye color. For each of four individuals (A–D), the actual hair and eye color are shown, along with the predicted results from genotyping the 24 SNPs. Reprinted with permission from Liu, F., et al., "Colorful DNA polymorphisms in humans," *Seminars in Cell & Developmental Biology* 24:562, 2013.

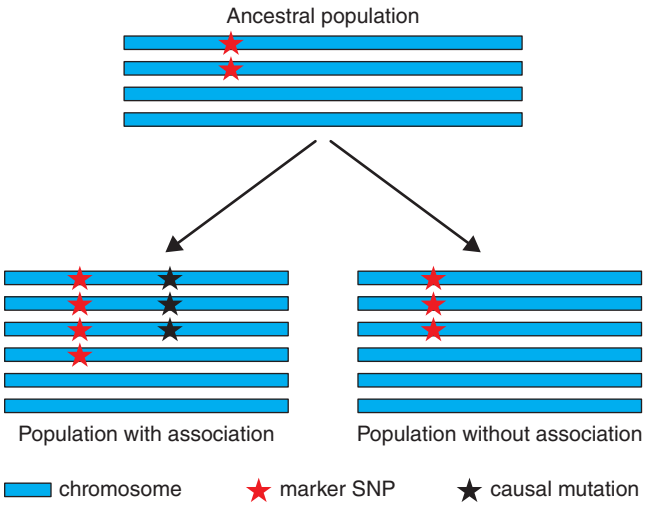


FIGURE 20.13

How a marker SNP may not be predictive of a causal mutation (black star) in every population. If we were to do a GWAS in the population on the left, we would find a strong association between the marker SNP and the trait in the population on the left, but the marker SNP and the trait would show no association in the population on the right.

unknown DNA sample by genotyping this associated mutation, you also need to know whether the person in question is of European or central Asian ancestry. Obviously, in this particular example, there would be no point in genotyping the associated mutation, since the causal mutation is known—but in many instances (such as risk of complex diseases, discussed later), we know only about associated mutations and not causal mutations.

So, genetics and molecular anthropology have contributed substantially to our knowledge concerning variation in skin/hair/eye pigmentation among individuals and populations. What about other phenotypic traits? Alas, with a few rare exceptions involving simple traits (e.g., wet vs. dry earwax, ability to taste certain bitter compounds, etc.), there hasn't been much to crow about. Take a trait like height, for example: easy to measure, has a high heritability of around 0.8 (meaning most of the variation among individuals reflects genetic differences rather than environmental differences in diet and the like) and varies substantially among individuals and populations. Of course, one expects there to be lots of different loci that influence height, but just how daunting the situation is only really became clear with a study by the aptly named GIANT (Genetic Investigation of Anthropocentric Traits) consortium (Lango Allen et al. 2010). This study was gigantic in many ways, as it was a genome-wide association study (GWAS) of 183,727 individuals, each genotyped at several hundred thousand SNPs, and, moreover, there are nearly 300 authors listed on the publication. The good news? A total of 180 loci were found to have significant effects on height (that's less than one per author!), including many novel genes and pathways, so there is much new biology to explore. The bad news? These 180 loci explain only about 10% of the variation in height among individuals, far less than the 80% actually expected to be contributed by genetic variation. This so-called "missing heritability" initially resulted in some consternation among researchers but is now generally recognized to reflect a combination of two factors: (1) there can be many different mutations in many different genes that influence height, only some of which will be detected in the GWAS approach; and (2) the alleles that influence height (and other complex traits) do not act independently but interact in complex ways (this is known as **epistasis**). A simple example of an epistatic interaction would be that if an individual has a mutation causing red hair and also another mutation causing albinism (complete loss of pigmentation), then the latter will completely mask the former—you won't know that the individual has the red hair mutation. However, the children of this individual might end up with red hair.

Anyway, the take-home message is that for height—and many other complex phenotypic traits of interest—what little we know about the genetics doesn't get us very far. As just one example, a study (Aulchenko et al. 2009) compared the predictive power of genotyping the 54 loci that showed the most significant association with height to the predictive power of a method published by Sir Francis Galton in 1886, which says that if you want to predict how tall someone will be, simply take the average of the heights of their parents (Galton 1886). The 54-locus genotypes accounted for just 4–6% of the variation in height, while Galton's method accounted for about 40% of the variation in height (Figure 20.14). So, Victorian methodology outperforms modern genomics by a factor of about 10! Clearly, there is much more to be done when it comes to figuring out—and making good use of—the genetic basis of complex phenotypic traits.

Another aspect of determining the relationship between genotypes and phenotypes was touched upon in Chapters 17 and 18, namely, how does one go about figuring out all of the various phenotypic effects that might be associated with a particular mutation of interest (e.g., one that shows a signature of selection)? As we saw with the examples involving FOXP2 and EDAR, there are a number of methods one can use, such as transcription assays, cell line assays, and even humanized mice. With luck, these sorts of studies might even point to particular phenotypes that then could be tested via association studies in humans (such as thicker hair associated with an EDAR mutation in East Asians).

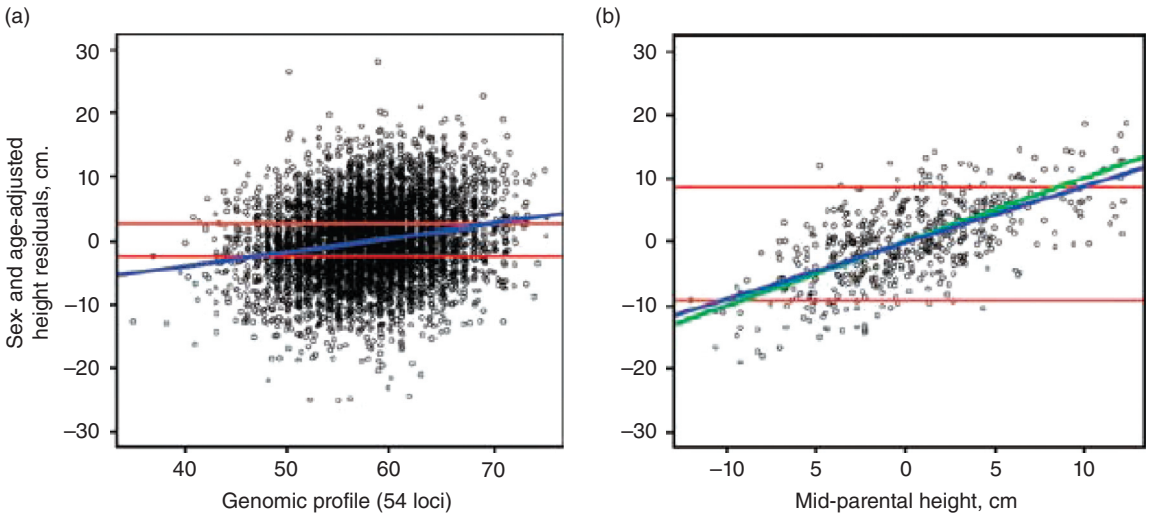


FIGURE 20.14

Correlation between actual height and that predicted by (a) genotypes for 5748 individuals based on 54 loci shown to be associated with variation in height, and (b) the average of the heights of the parents (midparental height) for 550 individuals. Red lines denote the top and bottom 5% of each distribution; the blue line is the best fitting linear regression line; the green line is the expected regression line assuming a perfect fit. Reprinted with permission from Aulchenko, Y.S., "Predicting human height by Victorian and genomic methods," *European Journal of Human Genetics* 17:1070, 2009.

But there remains much room for improvement, and one of the promising methods on the horizon involves a special type of cell called **induced pluripotent stem cells** (or iPSCs). Pluripotent means that these cells are capable of differentiating into many different types of specialized cells (neurons, heart cells, etc.), and the fact that they are stem cells means that they can remain in an undifferentiated state indefinitely while retaining the pluripotent ability to differentiate into various cell types. It used to be that the only pluripotent stem cells known were embryonic stem cells, which as the name suggests occur naturally during early embryonic development. With advances in cell culture techniques, embryonic stem cells attracted the interest of medical researchers when it was shown that they could be induced to form various types of cells that could be grown and studied in culture—neurons, for example, or even cardiomyocytes (heart cells) that started beating spontaneously in culture. But in order to harvest and study embryonic stem cells, the embryo must be destroyed, which in the case of human embryos raises all sorts of ethical issues.

It, therefore, was considered a major breakthrough when in 2006 the Japanese researcher Shinya Yamanaka and colleagues were able to transform skin cells from an adult mouse into pluripotent stem cells (thus, induced pluripotent stem cells or iPSCs) by manipulating four genes that are all involved in regulation of transcription (Takahashi and Yamanaka 2006). A year later, Yamanaka showed that human iPSCs could be obtained by a similar process (Takahashi et al. 2007), and for this work he was awarded a Nobel Prize in 2012. The dream is that if, in the future, you should find yourself in need of a new organ (say, your heart starts to fail), then instead of having to find a suitable donor, doctors could take some of your skin cells, induce them to become iPSCs, and then differentiate them into a new heart that could be transplanted into you without any worry of tissue rejection. At the moment this dream is far from reality—a major stumbling block is that the transcription factors that are manipulated when making iPSCs are also involved in cancer and tumor formation. You obviously don't want to transplant an iPSC-derived tissue if it's going to end up giving you cancer, so that all needs to be sorted out.

But there are other uses to which we might put iPSCs. In particular, say there is a mutation you are interested in that you think might alter how one particular organ functions—the heart, for example. With iPSCs, we could introduce the mutation of interest, induce the resulting modified iPSCs to differentiate into a heart, and then study various aspects of heart function, all in culture. To be sure, there is still a lot that we would like to (or need to) know about how a mutation functions within the context of an entire living, breathing individual—or group of individuals. Nevertheless, one can envision many situations in which iPSCs could provide key information that could not be obtained in any other way, and there are already experiments underway to investigate how best to go about using iPSCs in this context.

So to summarize this section, there are several very nice stories in which genetic approaches have contributed new insights into phenotypic variation (such as skin pigmentation variation). There are also some not so nice stories in which—despite extensive studies—genetic approaches have not contributed greatly to our understanding of phenotypic variation (such as height). Relating genotypes to phenotypes remains a very difficult but very essential problem in molecular anthropology, and the expectation (and hope) is that clever people will come up with new approaches for tackling this problem.

PERSONAL ANCESTRY TESTING AND GENOMICS

One aspect of being an anthropologist that you quickly learn to deal with is that not only does everyone seem to have an opinion about human origins and evolution, many do not hesitate to let you know what they think. I have received countless communications from all sorts of people from all walks of life, telling me how I should be doing my research. Judging from such correspondence, anthropology seems to be a favorite hobby of engineers in particular, which makes me wonder whether engineers similarly get letters from anthropologists telling them how they should be constructing their bridges or highways. Anyway, discounting the racists ("maybe your mother was black, but don't say that about my mother" was a common theme after the recent African mtDNA ancestor story came out in 1987) and the crazies (after the first Neandertal DNA was obtained in 1997, several people wrote to say that they knew people who were Neandertals that we should study) and the like, it is overall quite stimulating and encouraging to see the interest that people take in this work and the sorts of interpretations and questions that can arise.

One direction in which this popular fascination with human origins has led is using genetics to find out more about one's own ancestry, and beginning around 2000 a number of companies were quick to capitalize on this burgeoning interest by offering anyone the opportunity to find out about their genetic ancestry (at a suitable price, of course). Such personal ancestry testing got off to an inauspicious start when one of the first genetic ancestry companies, Oxford Ancestors, told customers that they would learn which of the "Seven Daughters of Eve" they were descended from. The idea they were trying to promote is that most Europeans belong to one of seven mtDNA haplogroups (H, J, K, T, U, V, X), which in turn are descended from the common African mtDNA ancestor, or "Eve" (via haplogroups M and N, which are derived from haplogroup L3, as we saw back in Chapter 9). Fair enough—but Oxford Ancestors gave names to each of these seven daughters, corresponding to the haplogroup names (Helena, Jasmine, Katrine, Tara, Ursula, Velda, and Xenia, respectively), and came up with a very imaginative (and, of course, completely imaginary) story about the lives of each of these women. So, for example, if you were haplogroup U, then you received a nice certificate telling you that you are descended from Ursula, a slender and graceful brunette who lived about 45,000 years ago in Greece and hunted bison with stone tools. And apparently, if you were so unlucky as to not have an mtDNA sequence from one of these seven mtDNA haplogroups, you didn't get a nice story about your particular daughter of Eve—some have referred to these other haplogroups as the "step-daughters of Eve"!. It seemed to many of us at the time that this sort of ancestry testing was little more than a joke, with much more emphasis on entertainment than on science.

Fortunately, ancestry testing has matured considerably. Nowadays, you can learn about your mtDNA (and, if you are male, Y chromosome) lineage and see where you fall in the worldwide phylogenetic tree, where else in the world your particular lineage(s) occurs (without resorting to completely fictitious stories about Ursula and the like), and even—if you are so inclined—make contact with others who share your lineage(s) to see whether you might be related. Most recently, ancestry testing has been extended to include genome-wide SNP typing, which potentially can tell you a lot more about your ancestry. Still, while the science has improved, one still has to be careful about the interpretations that some companies provide. For example, an ancestry test introduced a few years ago that was based on a few select ancestry informative markers gave many Europeans the disconcerting result that they had up to 12% Native American ancestry (corresponding to a great grandparent!). It is not so unusual for European–Americans to find that they have some Native American ancestry that they didn't know about, but it is rather more difficult to explain how people whose ancestors have always been living in Europe might have Native American ancestry. A little sleuthing revealed that these Europeans with supposed Native American ancestry actually have ancestry from central Asia, but the company that carried out the testing did not include any central Asians in their reference populations. You may remember from our discussion of admixture back in Chapter 12 that it is crucial in any analysis of admixture to have the right parental populations. In this case, in the absence of central Asian populations, the ancestry was assigned to the most closely related reference population that the company had data for, which happened to be Native Americans. More recently, there seems to be some issue with a new genome-wide ancestry test that promises to tell you how much of your ancestry

traces to Denisovans, judging from the number of e-mails I have received from people with "pure" European ancestry who nonetheless are told that they have around 4% Denisovan ancestry (which, you will recall from Chapter 16, we would expect only in people of Melanesian, Australian, or Filipino Negrito ancestry). Rather than some hidden Melanesian/Australian/Negrito ancestry in these people—or a previously undiscovered Denisovan contribution to Europeans—the most likely explanation is that the test is detecting something other than Denisovan ancestry (most probably, the test is confusing Denisovan with Neandertal ancestry).

Another aspect of personalized genome-wide tests that has attracted considerable attention is the information that you can learn about your risk of developing particular diseases. The leader in this field, a company called 23andMe, currently reports your risk of some 120 complex diseases, ranging from various cancers to diabetes to restless legs syndrome. However, there is a great deal of uncertainty surrounding such disease risk estimates. Very few of the alleles detected by the genome-wide tests actually cause the disease in question; instead, the vast majority of the alleles are only associated with the disease. That is, in a GWAS conducted in some population, people with the allele in question had a slightly higher risk of having the disease in question than people without the allele. And there's the rub—if you do not belong to the population in question, your disease risk based on having an associated allele may be quite different. Recall the example with eye color discussed previously in this chapter: the associated allele is highly predictive of blue eye color in Europeans but tells you nothing about eye color in central Asians. Even if you do belong to the population in question, the disease risk is the average risk associated with that allele; different people, with different genetic backgrounds and living in different environments, may very well have very different disease risks even though they all have the associated allele. As one commentator said a few years ago about the predictive value of genome studies for personal disease risk, all most of us can really expect to learn from our genomes about disease risk is that we should eat more sensibly, exercise more, and wear sunscreen when we go outdoors (Brenner 2007).

But let's be honest now—it is easy for us academics to sit in our ivory towers and smugly point out all the problems and potential errors associated with companies that are trying to earn a buck from telling people about their ancestry or their disease risk. After all, it's not like the academic literature is free from mistakes or erroneous conclusions and interpretations (far from it!). And all quibbles aside, personal ancestry testing has had an enormous influence on enhancing the public's interest in studies of genetic history, which to my mind at least, more than compensates for any errors of interpretation or questionable statements made for entertainment purposes. And even if personal disease risk estimates turn out to be quite wrong, still, if they get people to start thinking about what they can do to mitigate such risks and alter their behavior accordingly, then that's not such a bad thing, is it?

Moreover, the increasing interest in molecular anthropology (which is at least partly due to personal ancestry testing) has also increased public participation in molecular anthropology studies. As just one example, the Genographics Project was conceived in 2005 by the National Geographic Society (and the IBM Corporation) to pick up where the Human Genome Diversity Project left off, namely, by comprehensively surveying global genetic diversity in order to understand human history and migrations. The project had two components, one led by various researchers to sample particular geographic regions, the other based on public participation where you pay a fee to have your DNA analyzed—you get your results, but the results can also be used in the research. The research component has had variable success, as one might expect; some indigenous groups have refused to participate, and some researchers have been more successful than others at getting the necessary samples. But the public participation part has been a resounding success, with over half a million people participating to date, and some interesting studies have resulted, such as a study of nearly 80,000 mtDNA sequences (Behar et al. 2007). Other companies also utilize the results from their customers in research projects, and such research is becoming an increasingly important adjunct to molecular anthropology studies that rely on "traditional" sampling.

Finally, there is another form that public participation can take, and that is in actually carrying out molecular anthropology studies. To repeat what was stated at the end of Chapter 9, thanks to the ever-growing availability of public databases of DNA (and other molecular) data as well as the software to analyze such data, all you need is good Internet access and a reasonably fast computer to carry out your own research. Indeed, there is already an active community of amateur enthusiasts who maintain their own databases (all downloaded from public resources), maintain or even write their own software, and discuss in blogs and the like the results of various analyses. This book was written primarily with the beginning undergraduate student in mind, but it is my hope that anyone who has gotten this far in this book will take an interest in what can be learned about genetic history and be stimulated to do some investigating on his or her own. To paraphrase the last recorded words of Calvin (of Calvin and Hobbes fame), it's a magical history that we humans have, so let's go exploring!

REFERENCES

1000 Genomes Project Consortium, "A map of human genome variation from population scale sequencing," *Nature* 467:1061, 2010.

Abi-Rached, L., et al. "The shaping of modern human immune systems by multiregional admixture with archaic humans," *Science* 334:89, 2011.

Adovasio, J.M., and Carlisle, R.C., "The Meadowcroft rock shelter," *Science* 239:713, 1988.

Alexeyev, M., et al. "The maintenance of mitochondrial DNA integrity—critical analysis and update," *Cold Spring Harbor Perspectives in Biology* 5:a012641, 2013.

Allan, T.M., "Hirszfeld and the ABO blood groups," *British Journal of Preventive and Social Medicine* 17:166, 1963.

Allentoft, M.E., et al. "Population genomics of Bronze Age Eurasia," *Nature* 522:167, 2015.

Allison, A.C., "Protection afforded by sickle-cell trait against subtertian malarial infection," *British Medical Journal* 1:290, 1954a.

Allison, A.C., "The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria," *Transactions of the Royal Society of Tropical Medicine and Hygiene* 48:312, 1954b.

Allison, A.C., "The sickle cell and hemoglobin C genes in some African populations," *Annals of Human Genetics* 21:67, 1956.

Allison, A.C., "Malaria in carriers of the sickle cell trait and in newborn children," *Experimental Parasitology* 6:418, 1957.

Ammerman, A.J., and Cavalli-Sforza, L.L., *Neolithic Transition and the Genetics of Populations*, Princeton University Press: Princeton, NJ, 1984.

Anderson, S., et al. "Sequence and organization of the human mitochondrial genome," *Nature* 290:457, 1981.

Ankel-Simons, F., and Cummins, J.M., "Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution," *Proceedings of the National Academy of Sciences USA* 93:13859, 1996.

Aulchenko, Y.S., et al. "Predicting human height by Victorian and genomic methods," *European Journal of Human Genetics* 17:1070, 2009.

Avery, O.T., MacLeod, C.M., and McCarty, M., "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III," *Journal of Experimental Medicine* 79:137, 1944.

Axelsson, E., et al. "The genomic signature of dog domestication reveals adaptation to a starch-rich diet," *Nature* 495:360, 2013.

Balter, M., "Are humans still evolving?", *Science* 309:234, 2005.

Bamshad, M.J., et al. "Female gene flow stratifies Hindu castes," *Nature* 395:651, 1998.

Barbieri, C., et al. "Refining the Y chromosome phylogeny with southern African sequences," *Human Genetics* 135:541, 2016.

Bayes, T., and Price, R., "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S., communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S.," *Philosophical Transactions of the Royal Society* 53:370, 1763.

Beaumont, M.A., et al. "Approximate Bayesian computation in population genetics," *Genetics* 162:2025, 2002.

Begun, D.R. "Fossil record of Miocene hominoids," in W. Henke and I. Tattersall (editors), *Handbook of Paleoanthropology, Volume 2: Primate Evolution and Human Origins*, Springer: Berlin, p. 921, 2007.

Behar, D.M., et al. "The Genographic Project public participation mitochondrial DNA database," *PLoS Genetics* 3:e104, 2007.

Beleza, S., et al. "The timing of pigmentation lightening in Europeans," *Molecular Biology and Evolution* 30:24, 2013.

Berdoy, M., et al. "Fatal attraction in rats infected with Toxoplasma gondii," *Proceedings of the Royal Society B* 267:1591, 2000.

Berget, S.M., et al. "Spliced segments at the 5′ terminus of adenovirus 2 late mRNA," *Proceedings of the National Academy of Sciences USA* 74:3171, 1977.

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

Bernstein, F., "Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstructuren des Menschen," *Klinische Wochenschrift* 3:1495, 1924.

Bernstein, F., "Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen," *Zeitschrift für induktive Abstammungs- und Vererbungslehre* 37:237, 1925.

Bersaglieri, T., et al. "Genetic signatures of strong recent positive selection at the lactase gene," *American Journal of Human Genetics* 74:1111, 2004.

Bolnick, D.A., et al. "Asymmetric male and female genetic histories among native Americans from eastern North America," *Molecular Biology and Evolution* 23:2161, 2006.

Borradale, D.C., and Kimlin, M.G., "Folate degradation due to ultraviolet radiation: possible implications for human health and nutrition," *Nutrition Reviews* 70:414, 2012.

Bos, K.I., "A draft genome of *Yersinia pestis* from victims of the Black Death," *Nature* 478:506, 2011.

Bos, K.I., "Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis," *Nature* 514:494, 2014.

Bowcock, A.M., et al. "Study of 47 DNA markers in five populations from four continents," *Gene Geography* 1:47, 1987.

Bradley, B.J., et al. "Mountain gorilla tug-of-war: silverbacks have limited control over reproduction in multimale groups," *Proceedings of the National Academy of Sciences USA* 102:9418, 2005.

Bräuer, G., "The evolution of modern humans: a comparison of the African and non-African evidence," in P. Mellars and C.B. Stringer (editors), *The Origins and Dispersal of Modern Humans: Behavioural and Biological Perspectives*, Edinburgh University Press: Edinburgh, p.123,1989.

Brenner, S.E., "Common sense for our genomes," *Nature* 449:783, 2007.

Breton, G., et al. "Lactase persistence alleles reveal partial East African ancestry of southern African Khoe pastoralists," *Current Biology* 24:852, 2014.

Brown, W.M., "Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis," *Proceedings of the National Academy of Sciences USA* 77:3605, 1980.

Brown, W.M., et al. "Rapid evolution of animal mitochondrial DNA," *Proceedings of the National Academy of Sciences USA* 76:1967, 1979.

Bryk, J., et al. "Positive selection in east Asians for an EDAR allele that enhances NF-kappaB activation," *PLoS One* 3:e2209, 2008.

Buettner-Janusch, J., "The nature and future of physical anthropology," *Transactions of the New York Academy of Sciences* 31:128, 1969.

Cann, R.L., et al. "Mitochondrial DNA and human evolution," *Nature* 325:31, 1987.

Cano, R.J., et al. "Amplification and sequencing of DNA from a 120-135-million-year-old weevil," *Nature* 363:536, 1993.

Carneiro, M.O., et al. "Pacific biosciences sequencing technology for genotyping and variation discovery in human data," *BMC Genomics* 13:375, 2012.

Cavalli-Sforza, L.L., et al. *The History and Geography of Human Genes*, Princeton University Press: Princeton, NJ, 1994.

Cerqueira, C.C.S., et al. "Predicting *Homo* pigmentation phenotype through genomic data: from Neanderthal to James Watson," *American Journal of Human Biology* 24:705, 2012.

Chang, S.H., et al. "Enhanced EDAR signalling has pleiotropic effects on craniofacial and cutaneous glands," *PLoS One* 4:e7591, 2009.

Chaplin, G., and Jablonski, N.G., "Vitamin D and the evolution of human depigmentation," *American Journal of Physical Anthropology* 139:451, 2009.

Chargaff, E., et al. "The composition of the desoxyribonucleic acid of salmon sperm," *Journal of Biological Chemistry* 192:223, 1951.

Cheng, Z., et al. "A genome-wide comparison of recent chimpanzee and human segmental duplications," *Nature* 437:88, 2005.

Chikhi, L., et al. "Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool," *Proceedings of the National Academy of Sciences USA* 95:9053, 1998.

Chow, L.T., et al. "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA," *Cell* 12:1, 1977.

Coble, M.D., et al. "Mystery solved: the identification of the two missing Romanov children using DNA analysis," *PLoS One* 4:e4838, 2009.

Coehlo, M., et al. "Microsatellite variation and evolution of human lactase persistence," *Human Genetics* 117:329, 2005.

Collins, M.J., et al. "Is amino acid racemization a useful tool for screening for ancient DNA in bone?," *Proceedings of the Royal Society Biological Sciences* 276:2971, 2009.

Coon, C.S., *The Origins of Races*, Alfred A. Knopf: New York, NY, 1962.

Cooper, A., and Poinar, H.N., "Ancient DNA: do it right or not at all," *Science* 289:1139, 2000.

Cox, M.P., "Accuracy of molecular dating with the Rho statistic: deviations from coalescent expectations under a range of demographic models," *Human Biology* 81:911, 2009.

Cox, M.P., et al. "A Polynesian motif on the Y chromosome: population structure in remote Oceania," *Human Biology* 79:525, 2007.

Crick, F., "Central dogma of molecular biology," *Nature* 227:561, 1970.

Crick, F.H., "The origin of the genetic code," *Journal of Molecular Biology* 38:367, 1968.

Crow, J.F., "Thomas H. Jukes (1906-1999)," *Genetics* 154:955, 2000.

Darwin, G., "Note on the marriages of first cousins," *Journal of the Statistical Society* 38:344, 1875.

de Oliveira, F.B., et al. "Paleogeography of the South Atlantic: a route for primates and rodents into the New World?," in P.A. Garber, et al. (editors), *South American Primates: Comparative Perspectives in the Study of Behavior, Ecology, and Conservation*, Springer: Berlin, p. 55, 2009.

Diamond, J.M., "Archaeology: talk of cannibalism," *Nature* 407:25, 2000.

Dillehay, T.D., *Monte Verde: A Late Pleistocene Settlement in Chile" Paleoenvironment and Site Context.* Volume 1, Smithsonian Institution Press: Washington, DC, 1989.

Dillehay, T.D., et al. "New archaeological evidence for an early human presence at Monte Verde, Chile," *PLoS One* 10:e0141923, 2015.

Dorit, R., et al. "Absence of polymorphism at the *ZFY* locus on the human Y chromosome," *Science* 268:1183, 1995.
Drummond, A.J., et al. "Bayesian coalescent inference of past population dynamics from molecular sequences," *Molecular Biology and Evolution* 22:1185, 2005.
Duggan, A.T., and Stoneking, M., "A highly unstable recent mutation in human mtDNA," *American Journal of Human Genetics* 92:279, 2013.
Dunn, M., et al. "Structural phylogenetics and the reconstruction of ancient language history," *Science* 309:2072, 2005.
Eckhardt, R., "Stan Ulam, John von Neumann, and the Monte Carlo method," *Los Alamos Science* 15:131, 1987.
Enard, W., et al. "Molecular evolution of FOXP2, a gene involved in speech and language," *Nature* 418:869, 2002.
Enard, W., et al. "A humanized version of FOXP2 affects cortico-basal ganglia circuits in mice," *Cell* 137:961, 2009.
Enattah, N.S., et al. "Identification of a variant associated with adult-type hypolactasia," *Nature Genetics* 30:233, 2002.
ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature* 489:57, 2012.
Excoffier, L., et al. "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data," *Genetics* 131:479, 1992.
Falush, D., et al. "Traces of human migrations in *Helicobacter pylori* populations," *Science* 299:1582, 2003.
Fenner, J.N., "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies," *American Journal of Physical Anthropology* 128:415, 2005.
Feuk, L., et al. "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies," *PLoS Genetics* 1:e56, 2005.
Fischer, A., et al. "Demographic history and genetic differentiation in apes," *Current Biology* 16:1133, 2006.
Flink, L.G., et al. "Establishing the validity of domestication genes using DNA from ancient chickens," *Proceedings of the National Academy of Sciences USA* 111:6184, 2014.
Forster, P., et al. "Origin and evolution of Native American mtDNA variation: a reappraisal," *American Journal of Human Genetics* 59:935, 1996.
Fox, T.D., "Natural variation in the genetic code," *Annual Review of Genetics* 21:67, 1987.
Friedlaender, J.S., et al. "The genetic structure of Pacific Islanders," *PLoS Genetics* 4:e19, 2008.
Frisancho, A.R., and Baker, P.T., "Altitude and growth: a study of the patterns of physical growth of a high altitude Peruvian Quechua population," *American Journal of Physical Anthropology* 32:279, 1970.
Fu, Q., et al. "An early modern human from Romania with a recent Neanderthal ancestor," *Nature* 524:216, 2015.
Fujimoto, A., et al. "A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness," *Human Molecular Genetics* 17:835, 2008.
Galton, F., "Regression towards mediocrity in hereditary stature," *Journal of the Anthropological Institute* 15:246, 1886.
The Gene Ontology Consortium, "The Gene Ontology Project in 2008," *Nucleic Acids Research* 36:D440, 2008.
Gerbault, P., et al. "Evolution of lactase persistence: an example of human niche construction," *Philosophical Transactions of the Royal Society of London B Biological Sciences* 366:863, 2011.
Gilbert, M.T., et al. "DNA from pre-Clovis human coprolites in Oregon, North America," *Science* 320:786, 2008.
Giles, R.E., et al. "Maternal inheritance of human mitochondrial DNA," *Proceedings of the National Academy of Sciences USA* 77:6715, 1980.
Gill, P., et al. "Identification of the remains of the Romanov family by DNA analysis," *Nature Genetics* 6:130, 1994.
Gill, P., et al. "Establishing the identity of Anna Anderson Manahan," *Nature Genetics* 9:9, 1995.
Golenberg, E.M., "Chloroplast DNA sequence from a Miocene *Magnolia* species," *Nature* 344:656, 1990.
Gonçalves, V.F., et al. "Identification of Polynesian mtDNA haplogroups in remains of Botocudo Amerindians from Brazil," *Proceedings of the National Academy of Sciences USA* 110:6465, 2013.
Gongora, J., et al. "Indo-European and Asian origins for Chilean and Pacific chickens revealed by mtDNA," *Proceedings of the National Academy of Sciences USA* 105:10308, 2008.
Goodman, M., "Serological analysis of the systematics of recent hominoids," *Human Biology* 35:377, 1963.
Gray, R.D., et al. "Language phylogenies reveal expansion pulses and pauses in Pacific settlement," *Science* 323:479, 2009.
Green, R.E., et al. "A draft sequence of the Neandertal genome," *Science* 328:710, 2010.
Greenhill, S.J., et al. "How accurate and robust are the phylogenetic estimates of Austronesian language relationships?", *PLoS One* 5:e9573, 2010.
Grossman, S.R., et al. "A composite of multiple signals distinguishes causal variants in regions of positive selection," *Science* 327:883, 2010.
Gyllensten, U., et al. "Paternal inheritance of mitochondrial DNA in mice," *Nature* 352:255, 1991.
Gymrek, M., et al. "Identifying personal genomes by surname inference," *Science* 339:321, 2013.
Haak, W., et al. "Massive migration from the steppe was a source for Indo-European languages in Europe," *Nature* 522:207, 2015.
Hammer, M.F., et al. "Genetic evidence for archaic admixture in Africa," *Proceedings of the National Academy of Sciences USA* 108:15123, 2011.
Harding, R.M., et al. "Evidence for variable selective pressures at MC1R," *American Journal of Human Genetics* 66:1351, 2000.
Hardy, G.H., "Mendelian proportions in a mixed population," *Science* 28:49, 1908.
Harris, H., "Enzyme polymorphisms in man," *Proceedings of the Royal Society of London Series B Biological Sciences* 164:298, 1966.
Hawks, J., et al. "Recent acceleration of human adaptive evolution," *Proceedings of the National Academy of Sciences USA* 104:20753, 2007.
Hayes, B., "First link in the Markov chain," *American Scientist* 101:92, 2013.

Hedges, S.B., et al. "Human origins and analysis of mitochondrial DNA sequences," *Science* 255:737, 1992.

Hehemann, J.H., et al. "Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota," *Nature* 464:908, 2010.

Higuchi, R., et al. "DNA sequences from the quagga, an extinct member of the horse family," *Nature* 312:282, 1984.

Hsieh, P., et al. "Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies," *Genome Research* 26:291, 2016.

Hudson, R.R., et al. "A test of neutral molecular evolution based on nucleotide data," *Genetics* 116:153, 1987.

Huerta-Sánchez, E., et al. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA," *Nature* 512:194, 2014.

Hurles, M.E., et al. "Native American Y chromosomes in Polynesia: the genetic impact of the Polynesian slave trade," *American Journal of Human Genetics* 72:1282, 2003.

Hutchison, C.A., et al. "Maternal inheritance of mammalian mitochondrial DNA," *Nature* 251:536, 1974.

Ingram, C.J.E., et al. "Lactose digestion and the evolutionary genetics of lactase persistence," *Human Genetics* 124:579, 2009.

The International HapMap Consortium, "The International HapMap Project," *Nature* 426:789, 2003.

Jablonski, N.G., *Skin: A Natural History*, University of California Press: Berkeley, CA, 2006.

Jeffreys, A.J., et al. "Hypervariable 'minisatellite' regions in human DNA," *Nature* 314:67, 1985.

Johnson, M.J., et al. "Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns," *Journal of Molecular Evolution* 19:255, 1983.

Jordan, F.M., et al. "Matrilocal residence is ancestral in Austronesian societies," *Proceedings of the Royal Society Biological Sciences* 276:1957, 2009.

Jorde, L.B., et al. "Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data," *American Journal of Human Genetics* 57:523, 1995.

Jukes, T.H., and Cantor, C.R., "Evolution of protein molecules," in H.N. Munro (editor), *Mammalian Protein Metabolism*, Academic Press: New York, NY, p. 21, 1969.

Kamberov, Y.G., et al. "Modeling recent human evolution in mice by expression of a selected EDAR variant," *Cell* 152:691, 2013.

Kaneda, H., et al. "Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis," *Proceedings of the National Academy of Sciences USA* 92:4542, 1995.

Katzmarzyk, P.T., and Leonard, W.R., "Climatic influences on human body size and proportions: ecological adaptations and secular trends," *American Journal of Physical Anthropology* 106:483, 1998.

Kayser, M., et al. "Melanesian origin of Polynesian Y chromosomes," *Current Biology* 10:1237, 2000.

Kayser, M., et al. "Melanesian and Asian origin of Polynesians: mtDNA and Y chromosome gradients across the Pacific," *Molecular Biology and Evolution* 23:2234, 2006.

Khaitovich, P., et al. "A neutral model of transcriptome evolution," *PLoS Biology* 2:E132, 2004.

Kimura, M., "Evolutionary rate at the molecular level," *Nature* 217:624, 1968.

Kimura, M., "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *Journal of Molecular Evolution* 16:111, 1980.

Kimura, M., "*The Neutral Theory of Molecular Evolution*," Cambridge University Press: 1983.

Kimura, R., et al. "A common variation in EDAR is a genetic determinant of shovel-shaped incisors," *American Journal of Human Genetics* 85:528, 2009.

King, J.L., and Jukes, T.H., "Non-Darwinian evolution," *Science* 164:788, 1969.

King, M.C., and Wilson, A.C., "Evolution at two levels in humans and chimpanzees," *Science* 188:107, 1975.

Kirch, P.V., and Kahn, J.G., "Advances in Polynesian prehistory: a review and assessment of the past decade (1993-2004)," *Journal of Archaeological Research* 15:191, 2007.

Kittler, R., et al. "Molecular evolution of *Pediculus humanus* and the origin of clothing," *Current Biology* 13:1414, 2003.

Klein, R.J., et al. "Complement factor H polymorphism in age-related macular degeneration," *Science* 308:385, 2005.

Klopfstein, S., et al. "The fate of mutations surfing on the wave of a range expansion," *Molecular Biology and Evolution* 23:482, 2006.

Kocher, T.D., and Wilson, A.C., "Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region," in S. Osawa, and T. Honjo (editors), *Evolution of Life: Fossils, Molecules, and Culture*, Springer-Verlag: Tokyo, p. 391, 1991.

Kondo, R., et al. "Incomplete maternal transmission of mitochondrial DNA in Drosophila," *Genetics* 126:657, 1990.

Krause, J., et al. "The derived FOXP2 variant of modern humans was shared with Neandertals," *Current Biology* 17:1908, 2007.

Krause, J., et al. "The complete mitochondrial DNA genome of an unknown hominin from southern Siberia," *Nature* 464:894, 2010.

Krings, M., et al. "Neandertal DNA sequences and the origin of modern humans," *Cell* 90:19, 1997.

Kuhlwilm, M., et al. "Ancient gene flow from early modern humans into Eastern Neanderthals," *Nature* 530:429, 2016.

Kumar, S.S., et al. "Brief communication: discouraging prospects for ancient DNA from India," *American Journal of Physical Anthropology* 113:129, 2000.

Kumar, V., et al. "Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural process," *PLoS Genetics* 2:420, 2006.

Lahr, M.M., and Foley, R., "Multiple dispersals and modern human origins," *Evolutionary Anthropology* 3:48, 1994.

Lai, C.S., et al. "A forkhead-domain gene is mutated in a severe speech and language disorder," *Nature* 413:519, 2001.

Lalueza-Fox, C., et al. "A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals," *Science* 318:1453, 2007.

Lamason, R.L., et al. "SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans," *Science* 310:1782, 2005.

Landsteiner, K., "Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe," *Zentralblatt Bakteriologie* 27:357, 1900.

Landsteiner, K., and Wiener, A.S., "An agglutinable factor in human blood recognized by immune sera for rhesus blood," *Proceedings of the Society for Experimental Biology and Medicine* 43:223, 1940.

Langergraber, K.E., et al. "The genetic signature of sex-biased migration in patrilocal chimpanzees and humans," *PLoS One* 2:e973, 2007.

Langergraber, K.E., et al. "Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution," *Proceedings of the National Academy of Sciences USA* 109:15716, 2012.

Lango Allen, H., et al. "Hundreds of variants clustered in genomic loci and biological pathways affect human height," *Nature* 467:832, 2010.

Larson, G., et al. "Rethinking dog domestication by integrating genetics, archeology, and biogeography," *Proceedings of the National Academy of Sciences USA* 109:8878, 2012.

Lazaridis, I., et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans," *Nature* 513:409, 2014.

Lee, H.R., and Johnson, K.A., "Fidelity of the human mitochondrial DNA polymerase," *Journal of Biological Chemistry* 281:36236, 2006.

Leo, N.P., et al. "The head and body lice of humans are genetically distinct (Insecta: Phthiraptera, Pediculidae): evidence from double infestations," *Heredity* 95:34, 2005.

Lewontin, R.C., and Hubby, J.L., "A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*," *Genetics* 54:595, 1966.

Ley, R.E., "Obesity and the human microbiome," *Current Opinion in Gastroenterology* 26:5, 2010.

Li, H., and Durbin, R., "Inference of human population history from individual whole-genome sequences," *Nature* 475:493, 2011.

Li, W.H., and Tanimura, M., "The molecular clock runs more slowly in man than in apes and monkeys," *Nature* 326:93, 1987.

Linz, B., et al. "An African origin for the intimate association between humans and *Helicobacter pylori*," *Nature* 445:915, 2007.

Lippold, S., et al. "Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences," *Investigative Genetics* 5:13, 2014.

Lipson, M., et al. "Reconstructing Austronesian population history in island Southeast Asia," *Nature Communications* 19:4689, 2014.

Liu, R., et al. "Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection," *Cell* 86:367, 1996.

Long, J.C., and Kittles, R.A., "Human genetic diversity and the nonexistence of biological races," *Human Biology* 75:449, 2003.

Macaulay, V., et al. "Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes," *Science* 308:1034, 2005.

Macholdt, E., et al. "Tracing pastoralist migrations to southern Africa with lactase persistence alleles," *Current Biology* 24:875, 2014.

Mackintosh, J.A., "The antimicrobial properties of melanocytes, melanosomes, and melanin and the evolution of black skin," *Journal of Theoretical Biology* 211:101, 2001.

Malakoff, D., "Bayes offers a 'new' way to make sense of numbers," *Science* 286:1460, 1999.

Malaspinas, A.S., et al. "Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil," *Current Biology* 24:R1035, 2014.

Mantel, N., "The detection of disease clustering and a generalized regression approach," *Cancer Research* 27:209, 1967.

Maricic, T., et al. "A recent evolutionary change affects a regulatory element in the human FOXP2 gene," *Molecular Biology and Evolution* 30:844, 2013.

Marlar, R.A., et al. "Biochemical evidence of cannibalism at a prehistoric Puebloan site in southwestern Colorado," *Nature* 407:74, 2000.

Mathieson, I., et al., "Genome-wide patterns of selection in 230 ancient Eurasians," *Nature* 528:499, 2015.

Mayer, E.A., et al. "Gut microbes and the brain: paradigm shift in neuroscience," *Journal of Neuroscience* 34:15490, 2014.

McDonald, J.H., and Kreitman, M., "Adaptive protein evolution at the Adh locus in Drosophila," *Nature* 351:652, 1991.

McEvoy, B.P., et al. "Human population dispersal 'Out of Africa' estimated from linkage disequilibrium and allele frequencies of SNPs," *Genome Research* 21:821, 2011.

Mead, S., et al. "Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics," *Science* 300:640, 2003.

Meltzer, D.J., et al. "On the Pleistocene antiquity of Monte Verde, Southern Chile," *American Antiquity* 62:659, 1997.

Mendel, J.G., "Versuche über Pflanzenhybriden", Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV, Abhandlungen:3, 1865.

Mendez, F.L., et al. "An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree," *American Journal of Human Genetics* 92:454, 2013.

Metropolis, N., and Ulam, S., "The Monte Carlo method," *Journal of the American Statistical Association* 44:335, 1949.

Metropolis, N., et al. "Equations of state calculations by fast computing machines," *Journal of Chemical Physics* 21:1087, 1953.

Meyer, M., et al. "A high-coverage genome sequence from an archaic Denisovan individual," *Science* 338:222, 2012.

Meyer, M., et al. "A mitochondrial genome sequence of a hominin from Sima de los Huesos," *Nature* 505:403, 2014.

Meyer, M., et al. "Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins," *Nature* 531:504, 2016.

Miller, C.T., et al. "*cis*-regulatory changes in *Kit Ligand* expression and parallel evolution of pigmentation in sticklebacks and humans," *Cell* 131:1179, 2007.

Montagu, M.F.A., "Aleš Hrdlička, 1869-1943," *American Anthropologist* 46:113, 1944.

Mooder, K.P., et al. "Population affinities of Neolithic Siberians" a snapshot from prehistoric Lake Baikal," *American Journal of Physical Anthropology* 129:349, 2006.
Moodley, Y., et al. "The peopling of the Pacific from a bacterial perspective," *Science* 323:527, 2009.
Moreno-Estrada, A., "Reconstructing the population genetic history of the Caribbean," *PLoS Genetics* 9:e1003925, 2013.
Moreno-Mayar, J.V., et al. "Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans," *Current Biology* 24:2518, 2014.
Morton, N.E., "The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type," *American Journal of Human Genetics* 8:80, 1956.
Mou, C., et al. "Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form," *Human Mutation* 29:1405, 2008.
Myles, S., et al. "Identifying genes underlying skin pigmentation differences among human populations," *Human Genetics* 120:613, 2007.
Nasidze, I., and Stoneking, M., "Mitochondrial DNA variation and language replacements in the Caucasus," *Proceedings of the Royal Society Biological Sciences* 268:1197, 2001.
Nasidze, I., et al. "Testing hypotheses of language replacement in the Caucasus: evidence from the Y chromosome," *Human Genetics* 112:255, 2003.
Nasidze, I., et al. "Mitochondrial DNA and Y-chromosome variation in the Caucasus," *Annals of Human Genetics* 68:205, 2004.
Nass, M.M., and Nass, S., "Intramitochondrial fibers with DNA characteristics. I. Fixation and electron staining reactions," *Journal of Cell Biology* 19:593, 1963a.
Nass, M.M., and Nass, S., "Intramitochondrial fibers with DNA characteristics. II. Enzymatic and other hydrolytic treatments," *Journal of Cell Biology* 19:613, 1963b.
Neel, J.V., and Schull, W.J., "The effect of inbreeding on mortality and morbidity in two Japanese cities," *Proceedings of the National Academy of Sciences USA* 48:573, 1962.
Nei, M., and Roychoudhury, A.K., "Genetic relationship and evolution of human races," *Evolutionary Biology* 14:1, 1982.
Novembre, J., and Stephens, M., "Interpreting principal component analyses of spatial population genetic variation," *Nature Genetics* 40:646, 2008.
Ohashi, J., et al. "The impact of natural selection on an ABCC11 SNP determining earwax type," *Molecular Biology and Evolution* 28:849, 2011.
Ollivier, I., et al. "Jules-Sébastian-César Dumon D'Urville: On the Islands of the Great Ocean," *The Journal of Pacific History* 38:163, 2003.
Oota, H., et al. "Human mtDNA and Y-chromosome variation is correlated with matrilocal vs. patrilocal residence," *Nature Genetics* 29:20, 2001.
Oppenheimer, S.J., and Richards, M., "Polynesian origins: slow boat to Melanesia?," *Nature* 410:166, 2001.
Orlando, L., et al. "Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse," *Nature* 499:74, 2013.
Osawa, S., and Jukes, T.H., "Codon reassignment (codon capture) in evolution," *Journal of Molecular Evolution* 28:271, 1989.
Ottoni, C., et al. "Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics," *Molecular Biology and Evolution* 30:824, 2012.
Pääbo, S., "Molecular cloning of ancient Egyptian mummy DNA," *Nature* 314:644, 1985.
Palmer, M.S., et al. "Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease," *Nature* 352:340, 1991.
Parra, E.J., et al. "Estimating African American admixture proportions by use of population-specific alleles," *American Journal of Human Genetics* 63:1839, 1998.
Patterson, N., et al. "Genetic evidence for complex speciation of humans and chimpanzees," *Nature* 441:1103, 2006.
Patterson, N., et al. "Ancient admixture in human history," *Genetics* 192:1065, 2012.
Pauling, L., Letter to S. Leonard Wadler, August 15 1966. Quoted in G.P. Rédei, *Encyclopedia of Genetics, Genomics, Proteomics, and Informatics*, 3rd Edition, Springer: Berlin, p. 1918, 2008.
Pauling, L., et al. "Sickle cell anemia, a molecular disease," *Science* 110:543, 1949.
Pérez-Lezaun, A., et al. "Microsatellite variation and the differentiation of modern humans," *Human Genetics* 99:1, 1997.
Perry, G.H., et al. "Diet and the evolution of human amylase gene copy number," *Nature Genetics* 39:1256, 2007.
Pickrell, J.K., et al. "The genetic prehistory of southern Africa," *Nature Communications* 3:1143, 2012.
Pickrell, J.K., et al. "Ancient west Eurasian ancestry in southern and eastern Africa," *Proceedings of the National Academy of Sciences USA* 111:2632, 2014.
Pilbeam, D., "Notes on *Ramapithecus*, the earliest known hominid, and *Dryopithecus*," *American Journal of Physical Anthropology* 25:1, 1966.
Pinhasi, R., et al. "Optimal ancient DNA yields from the inner ear part of the human petrous bone," *PLoS One* 10:e129102, 2015.
Poinar, H.N., et al. "Amino acid racemization and the preservation of ancient DNA," *Science* 272:864, 1996.
Prado-Martinez, J., et al. "Great ape genetic diversity and population history," *Nature* 499:471, 2013.
Pritchard, J.K., et al. "Population growth of human Y chromosomes: a study of Y chromosome microsatellites," *Molecular Biology and Evolution* 16:1791, 1999.
Pritchard, J.K. et al. "Inference of population structure using multilocus genotype data," *Genetics* 155:945, 2000.
Prüfer, K., et al. "The complete genome sequence of a Neanderthal from the Altai Mountains," *Nature* 505:43, 2014.
Prugnolle, F. et al. "Geography predicts neutral genetic diversity of human populations," *Current Biology* 15:R159, 2005.
Prusiner, S., "Novel proteinaceous infectious particles cause scrapie," *Science* 216:136, 1982.
Ptak, S.E., et al. "Linkage disequilibrium extends across putative selected sites in FOXP2," *Molecular Biology and Evolution* 26:181, 2009.
Pugach, I., et al. "Genome-wide data substantiate Holocene gene flow from India to Australia," *Proceedings of the National Academy of Sciences USA* 110:1803, 2013.

Pugach, I., et al. "The complex admixture history and recent southern origins of Siberian populations," *Molecular Biology and Evolution* 33:1777, 2016.

Qin, P., and Stoneking, M., "Denisovan ancestry in East Eurasian and Native American populations," *Molecular Biology and Evolution* 32:2665, 2015.

Raff, J.A., et al. "Ancient DNA perspectives on American colonization and population history," *American Journal of Physical Anthropology* 146:503, 2011.

Rasmussen, M., et al. "Ancient human genome sequence of an extinct Palaeo-Eskimo," *Nature* 463:757, 2010.

Rasmussen, M., et al. "An Aboriginal Australian genome reveals separate human dispersals into Asia," *Science* 334:94, 2011.

Ray, N., et al. "Settlement of the American continent emphasizes the importance of gene flow with Asia," *Molecular Biology and Evolution* 27:337, 2010.

Reed, D.L., et al. "Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice," *BMC Biology* 5:7, 2007.

Reich, D., et al. "Reconstructing Indian population history," *Nature* 461:489, 2009.

Reich, D., et al. "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature* 468:1053, 2010.

Reich, D., et al. "Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania," *American Journal of Human Genetics* 89:516, 2011.

Reich, D., et al. "Reconstructing Native American population history," *Nature* 488:370, 2012.

Rieder, M.J., et al. "Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome," *Nucleic Acids Research* 26:967, 1998.

Risch, N., et al. "Categorization of humans in biomedical research: genes, race and disease," *Genome Biology* 3:comment 2007, 2002.

Roach, J.C., et al. "Analysis of genetic inheritance in a family quartet by whole-genome sequencing," *Science* 328:636, 2010.

Roberts, D.F., "Genetic effects of population size reduction," *Nature* 220:1084, 1968.

Robins, A.H., "The evolution of light skin color: role of vitamin D disputed," *American Journal of Physical Anthropology* 139:447, 2009.

Rogers, A.R., and Harpending, H., "Population growth makes waves in the distribution of pairwise genetic differences," *Molecular Biology and Evolution* 9:552, 1992.

Rogers, A.R., et al. "Genetic variation at the MC1R locus and the time since loss of human body hair," *Current Anthropology* 45:105, 2004.

Romeo, G., and Bittles, A.H., "Consanguinity in the contemporary world," *Human Heredity* 77:6, 2014.

Rosenberg, N.A., et al. "Genetic structure of human populations," *Science* 298:2381, 2002.

Roullier, C., et al. "Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination," *Proceedings of the National Academy of Sciences USA* 110:2205, 2013.

Sagan, L., "On the origin of mitosing cells," *Journal of Theoretical Biology* 14:225, 1967.

Saitou, N., and Nei, M., "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution* 4:406, 1987.

Sánchez-Quinto, F., et al. "Genomic affinities of two 7,000-year-old Iberian hunter-gatherers," *Current Biology* 22:1494, 2012.

Sanger, F., and Coulson, A.R., "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *Journal of Molecular Biology* 94:441, 1975.

Sanger, F., Nicklen, S., and Coulson, A.R., "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences USA* 74:5463, 1977.

Sankararaman, S., et al. "The genomic landscape of Neanderthal ancestry in present-day humans," *Nature* 507:354, 2014.

Sarich, V.M., "A molecular approach to the question of human origins," in P. Dolhinow and V.M. Sarich (editors), *Background for Man*, Little, Brown and Co.: Boston, p. 60, 1971.

Sarich, V.M., "Just how old is the hominid line?," *Yearbook of Physical Anthropology* 17:98, 1973.

Sarich, V.M., and Wilson, A.C., "Immunological time scale for hominid evolution," *Science* 158:1200, 1967a.

Sarich, V.M., and Wilson, A.C., "Rates of albumin evolution in primates," *Proceedings of the National Academy of Sciences USA* 58:142, 1967b.

Sawyer, S., et al. "Nuclear and mitochondrial DNA sequences from two Denisovan individuals," *Proceedings of the National Academy of Sciences USA* 112:15696, 2015.

Scally, A., and Durbin, R., "Revising the human mutation rate: implications for understanding human evolution," *Nature Reviews Genetics* 13:745, 2012.

Schatz, G., et al. "Deoxyribonucleic acid associated with yeast mitochondria," *Biochemical and Biophysical Research Communications* 15:127, 1964.

Schieffelin, E.L., and Crittenden, R., *Like People You See in a Dream: First Contact in Six Papuan Societies*, Stanford University Press: Stanford, CA, 1991.

Schiffels, S., and Durbin, R., "Inferring human population size and separation history from multiple genome sequences," *Nature Genetics* 46:919, 2014.

Schrago, C.G., et al. "Combining fossil and molecular data to date the diversification of New World primates," *Journal of Evolutionary Biology* 26:2438, 2013.

Schubert, M., et al. "Prehistoric genomes reveal the genetic foundation and cost of horse domestication," *Proceedings of the National Academy of Sciences USA* 111:E5661, 2014.

Schull, W.J., Otake, M., and Neel, J.V., "Genetic effects of the atomic bombs: a reappraisal," *Science* 213:1220, 1981.

Schurr, T.G., et al. "Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages," *American Journal of Human Genetics* 46:613, 1990.

Schwartz, C., "Trends and variation in assortative mating: causes and consequences," *Annual Review of Sociology* 39:451, 2013.

Schwartz, M., and Vissing, J., "Paternal inheritance of mitochondrial DNA," *New England Journal of Medicine* 347:576, 2002.

Seielstad, M., et al. "Genetic evidence for a higher female migration rate in humans," *Nature Genetics* 20:278, 1998.

Sharon, G., et al. "Commensal bacteria play a role in mating preference of *Drosophila melanogaster*," *Proceedings of the National Academy of Sciences USA* 107:20051, 2013.

Shimomura, Y., et al. "A rare case of hypohidrotic ectodermal dysplasia caused by compound heterozygous mutations in the EDAR gene," *Journal of Investigative Dermatology* 123:649, 2004.

Simonson, T.S., et al. "Genetic evidence for high-altitude adaptation in Tibet," *Science* 329:72, 2010.

Simoons, F.J., "Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations, II. A culture historical hypothesis," *American Journal of Digestive Diseases* 15:695, 1970.

Skoglund, P., et al. "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe," *Science* 336:466, 2012.

Smith, H.O., and Wilcox, K.W., "A restriction enzyme from Hemophilus influenzae. I. Purification and general properties," *Journal of Molecular Biology* 51:393, 1970.

Smithies, O., "Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults," *The Biochemical Journal* 61:629, 1955.

Soares, P., et al. "Ancient voyaging and Polynesian origins," *American Journal of Human Genetics* 88:239, 2011.

Soldevila, M., et al. "The prion protein gene in humans revisited: lessons from a worldwide resequencing study," *Genome Research* 16:231, 2006.

Southern, E.M., "Detection of specific sequences among DNA fragments separated by gel electrophoresis," *Journal of Molecular Biology* 98:503, 1975.

Steenbock, H., "The induction of growth promoting and calcifying properties in a ration by exposure to light," *Science* 60:224, 1924.

Steiper, M.E., et al. "Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence," *Proceedings of the National Academy of Sciences USA* 101:17021, 2004.

Stern, C., "The Hardy-Weinberg Law," *Science* 97:137, 1943.

Stoneking, M., "Women on the move," *Nature Genetics* 20:219, 1998.

Stoneking, M., "Widespread prehistoric human cannibalism: easier to swallow?," *Trends in Ecology and Evolution* 18:489, 2003.

Stoneking, M., et al. "Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa," *Genome Research* 7:1061, 1997.

Storey, A.A., et al. "Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile," *Proceedings of the National Academy of Sciences USA* 104:10335, 2007.

Stranger, B.E., et al. "Population genomics of human gene expression," *Nature Genetics* 39:1217, 2007.

Stringer, C.B., and Andrews, P., "Genetic and fossil evidence for the origin of modern humans," *Science* 239:1263, 1988.

Sturm, R.A., and Duffy, D.L., "Human pigmentation genes under environmental selection," *Genome Biology* 13:248, 2012.

Sudamant, P.H., et al. "An integrated map of structural variation in 2,504 human genomes," *Nature* 526:75, 2015.

Tajima, F., "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism," *Genetics* 123:585, 1989.

Takahashi, K., and Yamanaka, S., "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell* 126:663, 2006.

Takahashi, K., et al. "Induction of pluripotent stem cells from adult human fibroblasts by defined factors," *Cell* 131:861, 2007.

Tamura, K., and Nei, M., "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Molecular Biology and Evolution* 10:512, 1993.

Thomson, R., et al. "Recent common ancestry of human Y chromosomes: evidence from DNA sequence data," *Proceedings of the National Academy of Sciences USA* 97:7360, 2000.

Thomson, V.A., et al. "Using ancient DNA to study the origins and dispersal of ancestral Polynesian chickens across the Pacific," *Proceedings of the National Academy of Sciences USA* 111:4826, 2014.

Tishkoff, S.A., et al. "Convergent adaptation of human lactase persistence in Africa and Europe," *Nature Genetics* 39:31, 2007.

Tishkoff, S.A., et al. "The genetic structure and history of Africans and African Americans," *Science* 324:1035, 2009.

Torroni, A., et al. "Asian affinities and continental radiation of the four founding native American mtDNAs," *American Journal of Human Genetics* 53:563, 1993.

Underhill, P.A., et al. "Y chromosome sequence variation and the history of human populations," *Nature Genetics* 26:358, 2000.

Vernot, B., and Akey, J.M., "Resurrecting surviving Neandertal lineages from modern human genomes," *Science* 343:1017, 2014.

Vernot, B., et al. "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals," *Science* 352:235, 2016.

Vigilant, L., et al. "African populations and the evolution of human mitochondrial DNA," *Science* 253:1503, 1991.

Vigne, J.D., "Zooarchaeological aspect of the Neolithic diet transition in the Near East and Europe and their putative relationships with the Neolithic demographic transition," in J.P. Bocquet-Appel and O. Bar-Josef (editors), *The Neolithic Demographic Transition and its Consequences*, Springer: Dordrecht, The Netherlands, p. 179, 2008.

Wahlund, S., "Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet," *Hereditas* 11:65, 1928.

Wang, S., et al. "Genetic variation and population structure in native Americans," *PLoS Genetics* 3:e185, 2007.

Watson, J.D., and Crick, F.H.C., "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature* 171:737, 1953.

Weber, J.L., and May, P.E., "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction," *American Journal of Human Genetics* 44:388, 1989.

Weidenreich, F., "Facts and speculations concerning the origin of *Homo sapiens*," *American Anthropologist* 49:187, 1947.

Weinberg, W., "Über den Nachweis der Vererbung beim Menschen," *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* 64:368, 1908.

The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature* 447:661, 2007.

Wilder, J.A., et al. "Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males," *Nature Genetics* 36:1122, 2004.

Wilkins, J.F., "Unraveling male and female histories from human genetic data," *Current Opinion in Genetics and Development* 16:611, 2006.

Willerslev, E., et al. "Ancient biomolecules from deep ice cores reveal a forested southern Greenland," *Science* 317:111, 2007.

Williamson, S.H., et al. "Localizing recent adaptive evolution in the human genome," *PLoS Genetics* 3:e90, 2007.

Wirth, T., et al. "Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh," *Proceedings of the National Academy of Sciences USA* 101:4746, 2004.

Wollstein, A., et al. "Demographic history of Oceania inferred from genome-wide data," *Current Biology* 20:1983, 2010.

Wolpoff, M.H., et al. "A general theory of hominid evolution involving the fossil evidence from east Asia," in F.H. Smith and F. Spencer, *The Origins of Modern Humans*, Liss: New York, p. 411, 1984.

Wolpoff, M.H., et al. "Multiregional evolution: a world-wide source for modern human populations," in M.H. Nitecki and D.V. Nitecki (editors), *Origins of Anatomically Modern Humans*, Plenum Press: New York, p. 175, 1994.

Wood, E.T., et al. "Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes," *European Journal of Human Genetics* 13:867, 2005.

Woodward, S.R., et al. "DNA sequence from Cretaceous period bone fragments," *Science* 266:1229, 1994.

Workman, P.L., et al. "Selection, gene migration and polymorphic stability in a U. S. White and Negro population," *American Journal of Human Genetics* 15:429, 1963.

Wright, S., "Breeding structure of populations in relation to speciation," *American Naturalist* 74:232, 1940.

Wright, S., "Isolation by distance," *Genetics* 28:114, 1943.

Wyman, A.R., and White, R., "A highly polymorphic locus in human DNA," *Proceedings of the National Academy of Sciences USA* 77:6754, 1980.

Yi, X., et al. "Sequencing of 50 human exomes reveals adaptation to high altitude," *Science* 329:75, 2010.

Yoshiura, K., et al. "A SNP in the ABCC11 gene is the determinant of human earwax type," *Nature Genetics* 38:324, 2006.

Yu, N., et al. "Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1," *Molecular Biology and Evolution* 18:214, 2001.

Zegura, S., et al. "High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas," *Molecular Biology and Evolution* 21:164, 2004.

Zhai, W., et al. "An investigation of the statistical power of neutrality tests based on comparative and population genetic data," *Molecular Biology and Evolution* 26:273, 2009.

Zhang, W., et al. "Evaluation of genetic variation contributing to differences in gene expression between populations," *American Journal of Human Genetics* 82:631, 2008.

Zilber-Rosenberg, I., and Rosenberg, E., "Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution," *FEMS Microbiology Reviews* 32:723, 2008.

Zuckerkandl, E., and Pauling, L., "Molecular disease, evolution, and genic heterogeneity," in M. Kasha and B. Pullman (editors), *Horizons in Biochemistry*, Academic Press: New York, p. 189, 1962.

SUGGESTIONS FOR ADDITIONAL READING

In general, the Internet is a good source of additional information on all of the topics covered in this book—mind you, not all of it is worth the paper it's written on, so don't believe everything you find. Still, searching on some key terms can get you pretty far. The DNA Learning Center out of Cold Spring Harbor Laboratory (http://www.dnalc.org/) is a good place to start for more on the basics of genetics. And there are many good introductory-level texts on human genetics out there, such as:

The Cartoon Guide to Genetics (I kid you not!) by L. Gonick and M. Wheelis

Human Genetics: The Basics by R. Lewis

For those who want to delve more into molecular genetics, try

Human Molecular Genetics, by T. Strachan and A. Read

Molecular Biology of the Gene, various editions by various authors

Be forewarned, however, that these are not for the faint of heart; they are regularly used as textbooks in upper level or even graduate courses in genetics.

There are many good texts on population genetics that can be consulted for further information on the topics covered in Chapters 3, 4, and 5, including:

Principles of Population Genetics by D.L. Hartl and A.G. Clark

Genetics of Populations by P.W. Hedrick

For those looking for something a little easier, try *A Primer of Population Genetics* (D.L. Hartl) or *Human Population Genetics* (J.H. Relethford). And for the truly ambitious who are not put off by Kolmogorov backward equations and the like, there is the bible of the field, namely, *An Introduction to Population Genetics* (J.F. Crow and M. Kimura).

There's a lot more to molecular evolution than the very simple ideas presented in Chapter 6, and among the many good books available are:

Fundamentals of Molecular Evolution by D. Graur and W.H. Li

Molecular Evolution and Phylogenetics by S. Kumar and M. Nei)

For more about gene duplication, see the aptly named *Evolution by Gene Duplication* (S. Ohno). And for a comprehensive review of all things related to mobile DNA elements, see *Mobile DNA III* (edited by N.L. Craig, M. Chandler, M. Gellert, A.M. Lambowitz, P.A. Rice, and S.B. Sandmeyer).

For an overview and synthesis of what we have learned from classical markers that also brings in linguistic and archaeological evidence, the serious student of molecular anthropology should consult the mammoth compendium *The History and Geography of Human Genes* (L.L. Cavalli-Sforza, P. Menozzi, and A. Piazza); the not-so-serious student can opt for the lighter version, *Genes, Peoples, and Languages* (L.L. Cavalli-Sforza). The remaining topics covered in this book are also ably covered to various degrees by the only other real textbook in this field, namely, *Human Evolutionary Genetics* (M. Jobling, E. Hollox, T. Kivisild, and C. Tyler-Smith).

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

INDEX

An Introduction to Molecular Anthropology, First Edition. Mark Stoneking.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.