Scientists are using DNA to determine how the peoples of the world are related to one another.

Learning Objectives

- Explain how gene trees reconstruct the historical relationships among alleles within and between populations.
- Compare and contrast gene trees and species trees.
- Discuss how coalescence helps explain phylogenetic relationships between closely related species.
- Explain how the distribution of coalescent events can be used to detect historical changes in the size of a population.
- Describe the methods scientists use to construct phylogenetic trees.
- Discuss the kinds of evidence used to determine the origin of tetrapods, humans, and HIV.
- Discuss how the neutral theory of evolution is used to deduce the timing of evolutionary events and the history of natural selection.
- Explain how phylogenetic approaches can assist researchers in identifying disease-causing genes.
- Compare and contrast the evolution of genome size in bacteria and eukaryotes.

Sarah Tishkoff has been traveling from one end of Africa to the other for well over a decade. She took her first trip to Africa as a graduate student in genetics at Yale University, and she still returns now that she’s a professor at the University of Pennsylvania (Figure 8.1). She has bounced along cratered roads in Tanzania, and she has traveled aboard hand-cranked ferries in the jungles of Cameroon. On her journeys, Tishkoff carries syringes, vials, and centrifuges. Her expansive goal for all this travel is to create a genetic portrait of the 1 billion people who live in Africa. She and her colleagues still have a long way to go toward reaching that goal, but they have collected DNA from more than 7000 people from more than 100 ethnic groups.

Tishkoff hopes to learn many things from this genetic portrait she is putting together. She and her colleagues are beginning to identify alleles that make some Africans more vulnerable to certain diseases and resistant to others. But she also travels to Africa to understand history—not just the history of Africans, but the history of all humans.

Tishkoff and her colleagues are creating a detailed genealogy of the human race. Their research has been crucial to our
current understanding of how our species evolved. Thanks to the work of Tishkoff and other researchers, we now know that our species evolved in Africa for hundreds of thousands of years. Only after thousands of generations did a small group of Africans migrate out of the continent, interbreed with Neanderthals and other human populations that are now extinct, and ultimately spread across Asia, Europe, and the New World (Campbell and Tishkoff 2008; Nielsen et al. 2017; Chapter 17).

In Chapter 4, we saw how scientists use morphological traits to construct evolutionary trees. Until the 1990s, these data were the only kind of evidence readily available to evolutionary biologists seeking to understand phylogenies. But since then, researchers like Tishkoff have started using powerful computers and DNA-sequencing technology to unlock an extraordinary historical archive stored in the genomes of all living things. They’ve combined this insight with an understanding of how allele frequencies of a population can change over time through mechanisms of evolution such as drift and selection (Chapter 6 and Chapter 7).

As we shall see in this chapter, we can reconstruct the history of genes as well as the species that carried them. We can even detect episodes of genetic drift and natural selection that occurred millions of years ago.

### 8.1 Gene Trees

Conceptually, molecular phylogenies are similar to morphological ones of the sort we discussed in Chapter 4. In each case, evolutionary biologists compare homologous characters in a group of organisms to reconstruct their relationships. In a morphological phylogeny of birds and other dinosaurs, those characters might include the presence or absence of feathers (see Figure 4.31). In the case of molecular phylogenies, the characters might be the presence of adenine at a certain position in a certain gene. Biologists can use both morphological and molecular characters to identify clades. The common ancestry of humans and bats thus explains not only the homology of their bones but also the homology of many regions of their genomes.

But there are also some important differences in how scientists construct phylogenies from molecules versus morphology. To appreciate them, let’s take a closer look at how genetic variation arises in populations and how, over millions of years, speciation affects the distribution of that variation.

For this discussion, let’s consider the gene **BRCA1**, which we first discussed in Chapter 7. Located on chromosome 17, BRCA1 normally functions as a tumor-suppressor gene. Mutations that disrupt it can dramatically increase a woman’s risk of developing breast or ovarian cancer. Although the **BRCA1** gene is more than 126,000 base pairs long, it takes only a single mutation to a single nucleotide to create this risk. Scientists have identified hundreds of mutations in **BRCA1** that can lead to cancer, but here we’ll just consider one: a rare mutation that converts a single G nucleotide to T.

People who carry this particular G-to-T mutation can theoretically end up with it in one of two ways. It may have arisen in the egg or sperm that combined to produce their zygote. Or they may have inherited it from one of their parents. Their parent may have inherited the mutation in turn from one of their parents, and so on back through history. But at some point, the gene must have undergone the spontaneous mutation from G to T in one of their ancestors.
For example, the T variant could have arisen in an egg that was later successfully fertilized with a sperm carrying the G variant of BRCA1. The zygote now carried one benign G variant and one pathogenic T variant. If this zygote then developed into a woman, she would be at greater risk of breast and ovarian cancer. And if she had children, approximately half would inherit her T version of BRCA1. (Recall that thanks to segregation, only half of the eggs or sperm will carry this copy of the gene.) Her children who inherited the pathogenic T variant might then pass it down to some of their own children.

Figure 8.2A represents this process in a simplified form. Each circle in the column on the far left represents a BRCA1 allele in a population. (For simplicity’s sake, we’ll only consider ten alleles.) The columns to the right each represent BRCA1 alleles in the subsequent generations. From one generation to the next, some alleles are replicated, whereas others fail to be transmitted. The black lines highlight the descendants of a single allele. The orange line represents the transition from G to T in a single BRCA1 allele. In the final generation, on the right-hand side, there are now two copies of the pathogenic allele in the population.

The two orange circles and the black circle in the final generation are related to each other because they share common ancestors. The two orange circles are more closely related to each other than they are to the black circle, as evidenced by the G-to-T mutation that they share. If we want to represent their relationship, we can draw a genetic version of a family tree, akin to the pedigree we drew for Charles II of Spain in Chapter 6 (Figure 6.22). In that tree, each node and branch tip represented a human being. Now, to trace the spread of the pathogenic version of BRCA1, we can make these nodes and tips represent alleles (Figure 8.2B; Rosenberg and Nordborg 2002).

When a pathogenic T variant of BRCA1 arises in one person, it is initially very rare in a population, existing side by side with more common variants. As the new variant is passed down to descendants, it can become more common. In Chapter 6, we discussed how mechanisms such as selection and drift can drive variants to fixation or eliminate them from a population. In this chapter, we will look at this process from a genealogical perspective. This perspective applies not just to one pathogenic variant but to every allele of every gene we carry. Each allele has a history, a path of descent that we can trace backward through time. Figure 8.3 places the origin of the T variant in a broader population.

We can see from this figure that when G becomes T in one lineage, it becomes phylogenetically informative. In other words, it is a synapomorphy, because individuals with a T are similar to each other given they each inherited this character state from a recent common ancestor. In Chapter 4, we saw how evolutionary biologists use many morphological synapomorphies to determine the relationship between species and clades. The same holds true for gene trees. We have focused on a single site in the BRCA1 gene in this section, but there are many sites across the entire gene that are variable from person to person.

Each of these variable sites can be treated as a separate character in a phylogenetic analysis, and from these clues biologists can reconstruct the long-term evolutionary history of the gene. The branching path of descent with modification that describes the genealogy of a gene is called a gene tree.

Figure 8.4A shows a hypothetical gene tree for BRCA1, reconstructed from 20 copies of the gene sampled from the same population. Constructing gene trees for alleles sampled within a single population can reveal clues about the past—whether the
population went through a bottleneck event, for example, or whether the gene recently experienced directional or stabilizing selection. (Box 8.1).

But gene trees constructed from individuals from different populations or species can illuminate events that occurred much further back in time. For example, Figure 8.4B shows a gene tree constructed from 2622 base pairs of the BRCA1 gene sampled across 57 species of mammals (Fleming et al. 2003). The common ancestor of these mammals lived about 100 million years ago. When Melissa Fleming of Johns Hopkins University and her colleague created this gene tree, they could trace the rise of mutations over this vast period of time. They determined that some regions of the BRCA1 gene have accumulated very few mutations in that period, suggesting that they may be especially vulnerable to negative fitness effects. As we will see later, such surveys can help cancer biologists better understand how mutations to BRCA1 affect human health.

Figure 8.3 In this figure, we expand the BRCA1 gene tree in Figure 8.2B to show the relationship of ten sampled alleles at one base pair. The hypothetical cancer risk mutation is indicated as a change at this site from a G to T, occurring in one branch and creating a polymorphism that now coexists with the ancestral allele in the population. (Data from Rosenberg and Nordborg 2002)

Figure 8.4 A: A hypothetical gene tree for BRCA1, showing the historical relationships among 20 alleles sampled from a single population. Because diploid individuals carry two copies of each gene, and they each have their own unique history, studies like this focus on sampled alleles rather than individuals. Often, thousands of base pairs of sequence are collected for each sampled allele. When these sequences are aligned, polymorphisms such as the one depicted in Figure 8.3 provide informative characters for reconstructing their historical relationships. B: Alleles can also be sampled from different populations or species, revealing their phylogenies over much greater spans of time. This shows a gene tree for the BRCA1 gene in mammals descending from a common ancestor approximately 160 million years ago. (Data from Fleming et al. 2003)
Clues from Coalescence

Coalescent events can tell us a great deal about the history of a population. They can tell us if the population has experienced any major changes in size or if it has undergone natural selection (Nielsen and Slatkin 2013).

To see how we can do this, let’s move forward in time through a gene tree (Box Figure 8.1.1A), counting the splitting events as they occur. One branch splits into two, and these each split to produce four, and so on. The number of splitting events increases exponentially.

To examine coalescent events, on the other hand, we just read the tree in the reverse direction, beginning with the branch tips and working our way back to base of the tree. As tips coalesce into their most recent common ancestors, the number of branches decreases. We expect a constant rate of coalescence to lead to an exponential decrease in the number of coalescent events as we move farther back through time. Most alternative alleles should coalesce soon after we move back from the tips of the tree, and only a few should coalesce deeper back in time.

The rate at which branches coalesce depends on the size of the population. We saw in Chapter 6 that neutral alleles (continued)
Box 8.1  Clues from Coalescence (continued)

are either lost or fixed faster in small populations than in larger ones. If we look at a gene tree of a small population, we expect its genealogies to converge onto a common ancestor in fewer generations than in large populations. For any two copies of a gene sampled from a haploid population, the average time to coalescence should be $2N$ generations, where $N$ is the number of individuals in the population. In a diploid population, the time would be $4N$ generations.

We can predict the distribution of coalescent events for a population if we can estimate its population size and generate thousands of randomly generated tree topologies for the number of alleles in question (Box Figure 8.1.1A shows just one out of many possible topologies). This prediction serves as our null hypothesis. If actual gene trees violate these predictions, this can point to interesting properties of the population.

One way for a population to violate these predictions is to expand dramatically over time. If it started out very small, then it will have more accumulated coalescent events early in its history than we’d expect (Box Figure 8.1.1B). Shrinking populations will produce the opposite violation of the null hypothesis. Coalescent events will occur slower than predicted deep in the tree, and faster than expected near the tips (Box Figure 8.1.1C).

Key Concepts

- Alternative copies of a gene exist side by side within populations, and they each have a lineage that traces their history back through time.
- Gene trees reconstruct the historical relationships among alleles within and between populations.

8.2 Estimating the Age of an Allele

Now that we have followed these alleles forward through the generations, let’s travel back in time. When we march forward, we talk about nodes as splitting events. When we march backward, nodes are the points at which two lineages converge, or coalesce, into a single ancestral lineage. These events of coalescence occur at the most recent common ancestor of any two alleles. For closely related alleles like the branches depicted in Figure 8.2, we don’t have to trace back many generations before the alleles coalesce. For more distantly related alleles we might have to trace back the generations much further. As we saw in Chapter 6, some alternative alleles may persist side by side in populations for many thousands of generations. Gene trees like the one in Figure 8.4A can be used to estimate the times of coalescence for all possible pairwise combinations of the different alleles in a population.

Coalescence is the process in which the genealogy of any pair of homologous alleles merges in a common ancestor.

We can also detect even finer scale patterns. For example, a population that was large, experienced an acute bottleneck event, and then expanded once again will leave a mark of this history on the topology of its tree. It will have an unexpectedly high number of nodes clustered in the middle region of the tree.

Selection can cause observed gene trees to deviate from the null expectations, too. Strong positive selection, for example, can quickly pull an allele to high frequency. Reconstructing the gene tree for this gene would show coalescence times that are shallower—closer to the tips of the tree—than expected by chance. Balancing selection, on the other hand, like we observed with the A and S alleles of the β-globin locus in Chapter 6, can maintain multiple alleles within a population indefinitely. Coalescent events on this sort of gene tree should occur much deeper on the tree than expected.

Since changes in population size and selection can each generate similar patterns in the distribution of coalescent events, how can biologists distinguish between them? One way is to look at several different genes. Neutrally evolving genes should provide a clear picture of changes to overall population size—and they should all provide the same picture—whereas particular genes under specific types of selection should stand out.
The fact that all extant copies of a gene eventually coalesce does not mean that the population originally consisted of only a single individual with that ancestral allele. It just means that a particular individual’s allele was the one that, out of all the alleles present at that time, later became fixed in the population. If we were to examine the extant copies of another gene, they would probably coalesce in a different ancestral individual (and at a different time).

The timing of coalescence can vary a great deal. It depends on factors such as whether alleles are under selection or not. Positive selection can accelerate the rise in frequency in an allele, for example, shortening the time to fixation—and leading to a short coalescence. Two alleles that experience little selection may coexist for a longer period of time. The more generations in which multiple alleles persist side by side, the further back in time we need to travel before they converge on a common ancestral allele. Powerful statistical approaches enable scientists to use information, like the frequencies of the alleles and the size of the population, to make accurate estimates of the likely times to coalescence for alleles in particular genes (Box 8.1).

Consider, for example, mitochondrial DNA. All humans carry DNA in their mitochondria, which they inherit solely from their mothers, without meiotic recombination. Certain mitochondrial mutations are more common in certain ethnic groups. That pattern would suggest that they arose after the initial divergence of the lineages of alleles found in living humans. How far back in human ancestry must we go to find the common ancestral allele?

We can’t find the answer directly. We don’t have a complete record of all the mitochondrial DNA in every human alive today. Nor do we have all the ancestors of humans going back millions of years. But we can narrow the range of likely answers by sampling the current distribution of alleles and applying statistical models to them.

Alan Templeton, a population geneticist at Washington University in St. Louis, used one of these models to estimate the time to coalescence of human mitochondrial DNA. He found a range of 152,000 to 473,000 years. Figure 8.5 shows the results of his analysis of mitochondrial DNA as well as a number of autosomal genes. The time to coalescence varies tremendously, from tens of thousands of years to millions of years. (For the mathematics of coalescent theory, see Templeton 2006; Wakeley 2008.)

- It is possible to trace the genealogies of genes back through time, reconstructing when mutations generated new alleles and how these alleles subsequently spread.

8.3 Gene Trees and Species Trees

Sometimes a population will split into two or more reproductively isolated populations. As we’ll explore in more detail in Chapter 13, this isolation can eventually lead to the origin of new species. It can take thousands or even millions of years for isolated populations to become clearly delineated species. During that process of speciation, their alleles will be passed from generation to generation, accumulating mutations and forming branched genealogical lineages—gene trees—along the way. After
populations become reproductively isolated, they continue to accumulate changes, and their evolutionary trajectories will diverge. (See Box 8.2 for a discussion of how genes and proteins are named, and see Box 8.3 for a discussion of how genes in different species are related to one another.)

Gene trees often resemble the phylogeny of the species in which they are embedded. Over long periods of time, a species can acquire fixed alleles at many different loci, and collectively these differences can distinguish it from other species. If that species then gives rise to new species, the subsequent species will also inherit these loci. As a result, DNA can provide a clear signal of the phylogeny of these species, much like the morphological traits we discussed in Chapter 4.

But the history of genes is not always the same as the history of species. Here we will consider two important mechanisms that can create this mismatch.

**Introgression**

Sometimes individuals from one species interbreed with individuals from another. If the hybrid offspring survive and they end up mating with individuals from either of the original species, then gene copies from one of the species can be introduced into the genomes of the other. This introduction of genes through hybridization is known as **introgression**.

If these gene copies happen to carry beneficial variation, then they may be favored by selection and retained within the genomes of the recipient species. A gene tree reconstructed from these particular regions of the genome would appear more closely related to gene copies from the donor species than to other gene copies sampled from the recipient species.

As we’ll see in later chapters, introgression is proving to be a significant factor in molecular evolution. In Chapter 10, we’ll see how introgression has introduced an allele that’s allowing snowshoe hares to adapt to warmer climates. And in Chapter 17, we’ll see how Neanderthal DNA has been introgressed into the human genome.

**Incomplete Lineage Sorting**

Gene trees can sometimes fail to match species trees even without introgression. To understand why, we need to take a closer look at the fate of a gene as a population diverges. Initially, the population will have many alleles of the gene. And when the population splits, several alleles may be carried together into both of the resulting
species. If one of these lineages splits again, some of the same alleles will be carried through once more.

But eventually some alleles will be lost to genetic drift. If we then sample alleles of the gene from each of the daughter species, we might pluck a different one of our original alleles from each of them. Depending on which of the alleles happened to persist in each of the species, the gene tree we reconstruct might not reflect the actual branching history of the species.

Figure 8.6 illustrates how this mismatch can happen. Both panels of the illustration show three species with the same phylogenetic relationship. Embedded in each species tree is a gene tree that relates all alleles for a particular gene to each other.

In Figure 8.6A, alleles of the gene present in species 1 and species 2 diverged more recently from one another than either did from the allele present in species 3. The same relationship holds true for the species themselves. In Figure 8.6B, on the other hand, the coalescent time of the alleles reaches back before the divergence of the species. All three alleles were already present by the time of the first speciation event, and which copy ultimately persisted in each species was the result of drift. In this case, the alleles and the species have different histories, a situation called **incomplete lineage sorting**. The allele sampled from species 2 happens to be more closely related to the one we sampled from species 3.

If we were to rely on this one gene alone to determine the phylogeny of the three species, we’d end up concluding that 2 and 3 are closely related sister species when, in fact, they are not. And if we used the coalescence time of the gene to estimate the ages of the species, we would make another error because, in this case, the coalescent time is much older than the species themselves.

**Incomplete lineage sorting** occurs when a genetic polymorphism persists through several speciation events. When fixation of alternative alleles eventually occurs in the descendant species, the pattern of retention of alleles may yield a gene tree that differs from the true phylogeny of the species.

---

**Figure 8.6** Two phylogenies of three species are depicted by the thick branches. Within these three lineages, the dots and lines represent the alleles of a gene carried by individuals. If we sample the same gene in one individual from each species, we can trace back their genealogies until they coalesce in an ancestral allele. **A**: Here, the gene tree we derive matches the species tree. That is, the copies of the gene in species 1 and 2 share a more recent ancestry than either does with the gene in species 3. **B**: In this case, the relationship between the species is the same as in A, but the coalescent time is long compared to the time it took for the species to branch. As a result, alleles in the ancestral species were sorted in the descendant species to produce a pattern that is discordant with the species tree. If we were to rely on this particular gene to reconstruct the phylogeny of these species, we might conclude erroneously that species 2 and 3 have a closer common ancestor. (Data from Rosenberg and Nordborg 2002)
As a general rule, if the branch lengths of a species tree are much longer on average than the coalescent times of the genes being analyzed, scientists are unlikely to encounter incomplete lineage sorting. But it can be a problem in other phylogenies, such as ones depicting adaptive radiations (Chapter 14).

**A Molecular Phylogeny for Apes**

Charles Darwin recognized that humans share a close common ancestry with chimpanzees, gorillas, and orangutans. Since Darwin’s day, anatomical studies have confirmed that kinship. More recently, scientists also found that the DNA of apes provides even stronger support for Darwin’s original hypothesis. When they sequenced individual genes, researchers would produce gene trees that place apes closer to humans than any other species.

But these gene trees also produced some puzzling conflicts. Studying some genes, researchers concluded that chimpanzees were more closely related to humans than were gorillas or orangutans. But studies on other genes pointed to gorillas as our closest relatives. These studies, it’s now clear, were bedeviled by incomplete lineage sorting.

Thanks to advances in DNA-sequencing technology, scientists can now compare not just individual genes but entire genomes. In 2018 a team of scientists led by Evan Eichler at the University of Washington assembled reconstructions of genomes of humans, chimpanzees, gorillas, and orangutans (Kronenberg et al. 2018). Whereas earlier technologies could allow scientists to reconstruct genomes from only short
fragments a few hundred bases long, Eichler’s team read fragments that were thousands of base pairs long or longer. As a result, they could compare the genomes with far greater precision than before.

The researchers surveyed the single-nucleotide polymorphisms (SNPs) in the genomes of humans and other apes. To do so, they compared every corresponding 1000-base-pair segment in each species and determined how much each segment differed from species to species. Across the entire genome, they found that chimpanzees and humans differ by 1.27%. In terms of SNPs, we are 98.8% identical to chimpanzees. Gorillas are only slightly more divergent, differing on average by 1.61%. Orangutans, on the other hand, differ on average by 3.12%.

These percentages reflect the evolutionary history of the great apes. Using their new genomes, the researchers determined that the ancestors of orangutans branched off early in the history of the clade from our own ancestors. Chimpanzees and gorillas are more closely related, and as a result they differ by a smaller amount.

But when the researchers generated phylogenies based only on each 1000-base-pair segment, they sometimes ended up with different trees. As shown in Figure 8.7, 64.4% of their trees produced the clade with chimpanzees as our closest relatives. But

![Figure 8.7](image-url)

**Figure 8.7** A: In 2018, Evan Eichler and colleagues explored the phylogeny of humans and other great apes by comparing their genomes. They lined up each 1000-pb segment and estimated the divergence of the human segment with that of other species. The top figure is a hypothetical illustration of the results if there were no incomplete lineage sorting. The actual results, shown below, demonstrate incomplete lineage sorting among humans, chimpanzees, and gorillas. B: Eichler and his colleagues generated a phylogeny from each segment. Most produced the correct result (top), but small fractions produced incorrect ones due to incomplete lineage sorting. (Data from Kronenberg et al. 2018)
in 17.5% of the trees, gorillas were closest. And in 18.6% of the trees, gorillas and chimpanzees were closer to each other than either was to humans.

The interpretation that best explains these results is that orangutans split off from other great apes a long time ago, and their relationship to us is not masked by incomplete lineage sorting. By contrast, the ancestors of gorillas, chimpanzees, and humans split from each other over a relatively brief period of time. As a result, a substantial number of variants did not sort completely. Their evolutionary history is different from the evolutionary history of great ape species.

This evolutionary history explains why early studies on ape phylogeny, based on only single genes, produced confusing results. Later, when Eichler and his colleagues took into account all the available evidence, they could see that chimpanzees are our closest living relatives.

- Because alternative alleles sometimes persist side by side for a very long time, they may be passed down to daughter species in a fashion that does not reflect the actual branching history of the species (incomplete lineage sorting).
- Occasionally, gene copies from one species will be introduced into the genome of a second species (introgression).
- Both incomplete lineage sorting and introgression result in gene trees that differ from the true phylogeny of the species.
- Because gene trees occasionally have histories unlike those of the species that carry them, scientists often use information from many genes when they infer phylogenies from molecular data.

8.4 Methods of Molecular Phylogenetics

An understanding of how alleles arise and spread in populations allows us to use DNA sequence data to reconstruct phylogenies. The tips of these phylogenies may represent individual organisms, entire populations, or even species. Whichever is the case, the basic method is the same. Scientists compare the genetic sequences in a group of taxa and then determine the phylogeny that best explains how an ancestral sequence gave rise to the current patterns through mutation and fixation.

Molecular phylogenetics has some important limits. Evolutionary biologists can use morphological characters to reconstruct the phylogeny of a clade that includes both living and extinct species—such as the 375-million-year-old Tiktaalik (Figure 4.22). But fossils of Tiktaalik preserve only bone and no DNA. After an organism dies, its DNA degrades, and the useful information it stores gradually disappears. Ancient DNA rarely survives more than 100,000 years. Thus evolutionary biologists can compare only living and recently extinct species in molecular phylogenies—a tiny fraction of the diversity of life over the course of its 3.7-billion-year history.

Despite this limit, molecular phylogenetics has revolutionized evolutionary biology in the past few decades. DNA and protein sequences can potentially yield vast amounts of information for scientists seeking to reconstruct phylogenies. Each base in a segment of DNA is, in effect, a separate character that can have one of four different character states: A, C, G, or T. Each amino acid in a protein can take one of 20 states (Chapter 5). In addition, insertions and other mutations can serve as synapomorphies, helping to identify monophyletic groups. In a large-scale morphological analysis, evolutionary biologists might examine a hundred different characters in a group of species. In molecular analyses, they regularly examine thousands or even millions of characters. Even though a molecular phylogeny typically includes only extant species, the patterns found within and among taxa still yield important clues about the nature of the common ancestors of those species that lived hundreds of millions of years ago.
To take advantage of this rich trove of information, evolutionary biologists must contend with the special challenges posed by molecular evolution. We’ve already seen how species trees and gene trees don’t always match. Another source of error is homoplasy. As we saw in Chapter 4, separate lineages can independently arrive at the same character state. Morphological homoplasy—such as the hydrodynamic shape of dolphins and tuna (Chapter 1)—is often the product of mutations to many interacting genes. Molecular homoplasy, on the other hand, can evolve much more easily. Because each base in a segment of DNA can exist in only one of four states, the probability that separate lineages will independently arrive at the same character state can be high. It’s also possible for a site to mutate to a new nucleotide and then mutate again back to the original state. Instead of providing a stronger phylogenetic signal, such reversals erode it.

Thus any attempt to reconstruct a branching pattern for the past is likely to incorporate data with a jumble of signals. Some data provide true signals that accurately reflect the branching history of the group, and some provide false signals arising from homoplasy or reversals. Scientists use a number of analytical approaches to select the phylogeny that best approximates the actual history of a group.

The maximum parsimony method rests on the logic that we explored in Chapter 4’s discussion of morphological phylogenies, namely, that the simplest solution is also the most reasonable one. When scientists use parsimony methods, they examine the distribution of characters among taxa on a number of trees with different topologies. They calculate how often those characters would have changed if a candidate phylogeny is correct. The tree with the fewest number of character state changes is considered the most parsimonious (Kitching et al. 1998; Swofford 2002).

Homoplasy can present a misleading picture of the most parsimonious tree, but scientists can use statistical methods to reduce its effect. They give extra weight to informative portions of genomes and less weight to ones that are more prone to homoplasy. Exon regions of protein-coding genes, for instance, are useful for reconstructing phylogenies of distantly related species. That’s because they typically evolve very slowly when they are under strong purifying selection. The changes that do occur in exon regions are typically functional and are thus conserved for long periods of time. These types of slow substitutions can provide clear signals for constructing phylogenies because they reflect the patterns of ancestry of the lineages.

Introns and intergene regions of DNA are often effectively neutral with respect to selection. Such noncoding regions of DNA have more variable sequences—more information to use in building a tree—but also more homoplasy due to random convergence of base pairs. By giving less weight to rapidly changing characters such as noncoding bases, scientists can minimize the conflicting signal arising from homoplasy in the DNA (Williams and Fitch 1990; Maddison and Maddison 1992). Scientists may reverse these choices, however, if they are studying recently diverged lineages. Slow-evolving regions of DNA may experience too little change to be useful, whereas fast-evolving regions will offer more information.

When scientists use maximum parsimony methods, their analysis may present them with a single tree that’s the most parsimonious. Or they may end up with a group of trees that are equally parsimonious, which can be combined into a single consensus tree that includes only branching patterns that appear in all of the most-parsimonious trees. In either case, scientists can then analyze the final tree with statistical methods to test how strongly the available data support it.

One of these tests is called bootstrapping. In this process, researchers select a random sample of characters from their full data set, much like drawing characters from a hat—with the caveat that each time a character is pulled it is immediately replaced in the hat, so that some characters may be sampled multiple times and others not at all. Scientists keep sampling the characters until they reach the same number as in the original set. Then they create a new data matrix and use it to generate a potential phylogeny. They repeat the process, randomly selecting characters and creating another phylogeny. After generating thousands of these potential trees, they can then compare them to each other. If the trees are very different from each other, it means

Maximum parsimony is a statistical method for reconstructing phylogenies that identifies the tree topology that minimizes the total amount of change, or the number of steps, required to fit the data to the tree.

Purifying selection (also called negative selection) removes deleterious alleles from a population. It is a common form of stabilizing selection (Chapter 7).

Bootstrapping is a statistical method that allows for assigning measures of accuracy to sample estimates; scientists use it for estimating the strength of evidence that a particular node in a phylogeny exists.
A distance-matrix method is a procedure for constructing phylogenetic trees by clustering taxa based on the proximity (or distance) between their DNA or protein sequences. These methods place closely related sequences under the same internal branch, and they estimate branch lengths from the observed distances between sequences.

Neighbor joining is a distance method for reconstructing phylogenies. Neighbor joining identifies the tree topology with the shortest possible branch lengths given the data.

Maximum likelihood is an approach used to estimate parameter values for a statistical model. Maximum likelihood and the similar Bayesian method are used in phylogeny reconstruction to find the tree topologies that are most likely, given a precise model for molecular evolution and a particular data set.

Bayesian methods refer to tests that are similar to maximum likelihood; they use statistical models to determine the probability of data given an evolutionary model and a phylogenetic tree.

the data offer poor support for the original tree. On the other hand, if the bootstrap test reveals that the trees are all very similar, it indicates stronger support.

Scientists can use bootstrapping to evaluate how reliable the support is for each branching event (node) in an evolutionary tree. They will often print the bootstrap values above each of the nodes in a phylogeny so that the relative strength of support for each part of the tree can be evaluated.

Other methods for generating phylogenetic trees don’t rely on the assumption of parsimony. For example, they may group taxa based on how different they are from each other. These distance-matrix methods convert DNA or protein sequences from different taxa into a pairwise matrix of the evolutionary distances (dissimilarities) between them. These methods predict that closely related species will have more similarities than more distantly related species, and this information can be used to group species into clades. These methods also can be used to estimate the lengths of the branches in the tree by equating the genetic distance between nodes with the length of the branch.

For example, neighbor joining is a distance-matrix method in which scientists pair together the two least-distant species by joining their branches at a node (Figure 8.8). They then join this node to the next-closest sequence, and so on. By joining neighboring species in round after round of this procedure, scientists can find the tree with the smallest possible distances—and the shortest possible branch lengths—between species (Saitou and Nei 1987).

One big advantage to using molecular data for phylogenetic studies is that we actually know a lot about the way DNA and protein sequences evolve (Chapter 6). For example, we’ve already seen that noncoding regions of DNA are likely to accumulate mutations faster than coding regions do. Within coding regions, substitutions that don’t affect the amino acid coded for by a site (Chapter 5) are likely to accumulate more rapidly than substitutions that do. This means that not all base pairs in a DNA sequence are expected to change at the same rate or in the same way. It also means that there’s a lot of additional information lurking in the sequences of species.

Maximum likelihood methods can use this extra information, and as a result they have become some of the most powerful and prevalent approaches for inferring phylogenies (Felsenstein 1981, 2004; Huelsenbeck and Crandall 1997). Maximum likelihood methods begin with an explicit model of evolution at the molecular level.

These models force some substitution rates to be equal. For example, the rate from A to C and from A to T might be set to have the same value. But the models allow other rates to vary. For example, the rate at one gene might be allowed to differ from the rate at another gene. Allowing less restricted models improves the realism of the model but comes at the cost of more difficulty in estimating the parameters.

These models of molecular evolution can be very complex, and they can be tailored to particular genes or regions of the sequences included. For example, one model can be used for coding regions and another model for noncoding regions. Scientists can also optimize these models to best explain the data they’re studying (for example, by incorporating mutation rates that are estimated from their data for the particular species being studied).

Once scientists have specified their model for molecular evolution, they can use it to calculate the probability of observing their data set, given the tree and the model. This result is known as the “likelihood.” In this approach to testing molecular phylogenies, the better trees are those for which the data are most probable. By comparing these likelihood scores for many different possible trees, scientists can identify the tree that best fits their data and their model for molecular evolution.

Another group of tests is known as Bayesian methods (Yang and Rannala 1997; Huelsenbeck and Ronquist 2001). They are similar to maximum likelihood in that they also use statistical models of the way DNA or protein sequences evolve. Maximum likelihood methods determine the probability of the data, given an evolutionary model and a hypothetical tree. Bayesian methods, on the other hand, determine the probability of a tree, given an evolutionary model and a particular data set.
Scientists start with a possible tree and then make small changes to its topology. They evaluate the new topology, given the data and the model, and then change the tree again. They repeat this procedure thousands of times and in the process generate a probability distribution for the different possible trees. Eventually, they converge on a set of the most likely trees. Rather than returning a single best tree and set of parameters, Bayesian methods seek to estimate the probabilities of a wide range of trees and parameters and thus give an estimate of the most probable history as well as the uncertainty of the results.

Over the past two decades, scientists have argued vigorously about the merits of each method we’ve presented here. What scientists agree on is that no single method for inferring molecular phylogenies is superior to all the rest. Although maximum likelihood methods have proven to be very powerful, for example, they also demand far more computer processing than, say, neighbor joining. In practice, scientists compare the results of several of these methods to explore the relationships of species and populations (Figure 8.9).
Figure 8.9 This phylogeny illustrates how biologists use maximum likelihood and Bayesian methods. It is the most likely tree topology based on comparison of 3051 base pairs of mitochondrial DNA sequence for Central American populations of poison dart frogs. Branch lengths in this figure reflect relative amounts of molecular evolution and can be used as an approximation for elapsed time between speciation events. The numbers associated with each node are statistical measures of support for its grouping of taxa. The blue number is the bootstrap support calculated from maximum likelihood models. The green number is the posterior probability calculated with Bayesian analysis. Together, these approaches reveal which portions of the phylogenetic hypothesis have the greatest (and weakest) support, providing a robust metric for evaluating the extent to which the data support a given tree. (Data from Wang and Schaffer 2008)
8.5 Three Case Studies in Molecular Phylogeny

As we saw in Chapter 4, scientists can use morphological characters to reconstruct phylogenies to test hypotheses about major transitions in evolution. Molecular phylogenetics also allows scientists to address these evolutionary questions. In this section we will look at three case studies, stretching across a huge range of taxa, that illustrate the power of this approach.

The Origin of Tetrapods

Molecular phylogeny enables scientists to reevaluate phylogenies that were developed based on morphology alone. Paleontologists, for example, have long argued that the closest living relatives of tetrapods were lobe-finned fishes, a group that today includes only lungfishes and coelacanths (Section 4.5). That was a fairly precise prediction, because there are around 30,000 species of fishes alive today. Of all those fishes, paleontologists predicted that only a half dozen should share a close common ancestry with tetrapods.

In 2017, Naoko Takezaki of Kagawa University and Hidenori Nishihara of the Tokyo Institute of Technology carried out a study of vertebrate phylogeny to test this hypothesis (Takezaki and Nishihara 2017). They compared DNA of a variety of tetrapods to that of lungfish and coelacanths. As an outgroup, they examined gars, which are a primitive lineage of ray-finned fishes. The phylogeny they reconstructed is shown in Figure 8.10. Their analysis shows lungfishes as the closest relative to tetrapods and coelacanths as the next-closest sister clade to tetrapods and lungfishes. Thus scientists studying one line of evidence—DNA—have confirmed a hypothesis originally developed from another line of evidence, the anatomy of fossil and living species.

How Did Homo Sapiens Evolve?

Earlier in this chapter, we saw how molecular phylogenetics illuminates our common ancestry with apes. As we’ll discuss in more detail in Chapter 17, fossil and molecular evidence suggests that the last common ancestor of humans and chimpanzees lived about 6 million years ago. Over the past 6 million years, our lineage has produced perhaps 20 different species, known collectively as hominins. One of the most important questions about hominin evolution that scientists have tried to answer has to do with our own species: how did *Homo sapiens* evolve?

Until the 1980s, the evidence that scientists could use to address that question was almost entirely limited to morphology. Based on the fossil record, a number of paleoanthropologists concluded that *Homo sapiens* had evolved gradually across the entire Old World from an older hominin species over the past 1 million years. This hypothesis came to be known as the “multiregional model” of human evolution.

In the 1980s, Chris Stringer of the National History Museum in London and other paleoanthropologists put forward a competing model. They argued that *Homo sapiens* evolved in Africa alone, and that other hominin fossils from the past million years were extinct branches. For evidence, they pointed to the fact that the earliest fossils with clearly modern anatomical traits are found in Africa, dating back as far as 300,000 years (Section 3.12). The oldest clearly modern fossils outside of Africa, found in Israel, date back only 100,000 years, and the fossil record of modern humans outside...
of Africa becomes strong only 50,000 years ago. Stringer and his colleagues proposed that *Homo sapiens* evolved in Africa and that some populations expanded to other continents much later (Stringer 2012).

This hypothesis generates a clear prediction: all major ethnic groups of humans—Africans, Europeans, and Asians—are derived from recent African ancestry. Once geneticists began gathering DNA from a wide range of human populations, they began to put that hypothesis to the test.

Sarah Tishkoff has been in a particularly good position to do this research because she has gathered so much genetic information about people in Africa, where Stringer and others proposed humans originated. Tishkoff and her colleagues analyzed DNA from Africans and compared their genetic sequences with those of people from other parts of the world.

The results of one study are shown in Figure 8.11 (Tishkoff et al. 2009). Tishkoff and her colleagues studied 121 African populations, 4 African American populations, and 60 non–African populations. They identified patterns of variation at 1327 genetic loci. Some of these loci, known as nuclear microsatellites, are stretches of repeating DNA that have a very high mutation rate. The researchers also looked at loci where DNA had been either inserted or deleted. Using the neighbor-joining method,
By comparing DNA from many populations of humans, Sarah Tishkoff and her colleagues used a neighbor-joining approach to produce an evolutionary tree of our species. Because humans have been in Africa much longer than in other parts of the world, Africans today are also much more genetically diverse than other humans. A small group of Africans migrated out of the continent and became the ancestors of today's Europeans, Asians, and people of the New World. (Data from Tishkoff et al. 2009)
they reconstructed a tree that revealed where most human genetic diversity can be found—and it is in Africa. What’s more, all non–Africans form a monophyletic group, suggesting that they diversified after migrating out of Africa.

Tishkoff’s results support the work of other researchers who have studied other genes (for example, whole mitochondrial genomes; Ingman et al. 2000). It appears that our species first evolved in Africa. Tishkoff finds the greatest level of diversity and the deepest branches among the people of Africa. Thousands of generations passed before some humans left the continent. Tishkoff’s research even offers hints about where in Africa they departed from. Her phylogeny suggests that when humans left Africa, they emigrated from East Africa—the same region where the oldest fossils of humans have been found.

In Chapter 17, we’ll look at other studies of human molecular phylogenetics that have enriched our understanding of our evolution even more. It turns out that Neanderthals and other extinct lineages of hominins have endowed us with a genetic legacy, having interbred with our ancestors.

**Determining the Origins of the Human Immunodeficiency Virus**

Molecular phylogenetics has become extremely important in the search for the origins of diseases. Today, for example, HIV is all too familiar. In 2017, an estimated 36.9 million people worldwide had HIV infections, and an estimated 940,000 people died of AIDS-related causes.

Yet, as diseases go, HIV is a latecomer. Scientists first became aware of it in the early 1980s, when it was still relatively rare. Soon after, HIV swiftly became a global epidemic. Scientists have searched through medical records and blood samples for earlier cases of HIV infection that might have been overlooked. The earliest known case of HIV comes from a blood sample taken from a patient in 1959 in what is now Kinshasa, the capital of the Democratic Republic of the Congo.

The mysterious appearance of HIV led to much speculation about where it came from. But when scientists analyzed its genetic material, a clear picture of its origins began to emerge.

HIV belongs to a group known as the lentiviruses. Lentiviruses infect mammals such as cats, horses, and primates, typically by invading certain types of white blood cells. Molecular phylogenetic studies revealed that HIV is most closely related to strains of lentivirus that infect monkeys and apes—collectively, these strains are known as simian immunodeficiency virus, or SIV for short. However, HIV is not a monophyletic clade. Instead, different strains of HIV have different origins (Van Heuverswyn et al. 2007; Sharp and Hahn 2011).

The virus known as HIV-1, which causes the vast majority of AIDS cases, is most closely related to the SIV viruses that infect chimpanzees. HIV-2 belongs to a group of SIV viruses that infect a monkey known as the sooty mangabey. A closer look at HIV-1 (Figure 8.12) reveals that it initially evolved in a subspecies of chimpanzee, *Pan troglodytes troglodytes*, found in Central Africa (d’Arc et al. 2015). Today, HIV-1 is classified into four groups, known as M, N, O, and P. Each group independently evolved from SIV in *P. t. troglodytes*. HIV-1 Group M and Group N both evolved directly from chimpanzee viruses. However, Group O and Group P had a more complex evolution. SIV jumped from chimpanzees to gorillas. Then the gorilla-adapted SIV jumped twice to humans. It’s likely that HIV evolved from apes and monkeys through hunting, as people came into contact with infected animals.

Knowing the structure of the HIV tree allows scientists to pinpoint the specific adaptations that may have allowed HIV to infect humans. It turns out, for example, that as HIV-1 lineages adapted to humans as a new host, they acquired the same mutation encoding the same new amino acid in the same position in the same protein (Wain et al. 2007). No SIV virus in chimpanzees codes for that amino acid in that position, and the phylogeny suggests that it arose independently each time the virus
adapted to its new human hosts. This mutation altered a gene encoding the shell of the virus, and experiments suggest that the mutation was crucial to the success of the new HIV viruses in humans. It’s possible that the mutation allowed the virus to better manipulate its hosts into building new copies of itself.

Studies like these allow scientists to better understand the evolution of human disease and may help to better predict the emergence of new pathogens. Molecular phylogenies have also helped solve legal cases involving reckless and criminally negligent transmission of HIV (Box 8.4), and they’ve played a pivotal role in detecting the origins of a recent outbreak of Ebola virus (Gire et al. 2014). We’ll explore the relevance of evolution to medicine in greater detail in Chapter 18.

- Constructing phylogenies is often a process of evaluating evidence. Scientists can test the predictions of phylogenetic hypotheses developed with one line of evidence by using other, independent lines of evidence to draw conclusions. •
8.6 Natural Selection Versus Neutral Evolution

In the last section, we saw how molecular phylogenies help scientists solve specific mysteries about particular taxa. They can also help scientists answer broad questions about the process of evolution. One of the biggest questions is how much of life’s diversity can be explained by evolutionary mechanisms other than selection.

Evolutionary biologists agree that natural selection is critically important to the evolution of complex morphology and behavior, because these phenotypes directly

Forensic Phylogenies

In 2005, a Texas man named Philippe Padieu learned from his doctor that he was infected with HIV. Despite the news, Padieu went on to have unprotected sex with a number of women, who later contracted HIV. After learning of Padieu’s reckless behavior, the local district attorney charged him in 2007 with six indictments for the offense of aggravated assault with a deadly weapon.

The prosecutors marshaled many lines of evidence to show that Padieu had knowingly infected his partners with HIV. And some of the most compelling evidence came from evolutionary biology.

When a pathogen such as HIV infects an individual (subject 1), it rapidly reproduces within him or her. In each of the lineages, distinctive mutations accumulate. If subject 1 then infects someone else, the pathogens in subject 2 will carry the mutations that arose in subject 1.

Scientists can discover this trail of genealogical evidence by comparing the genetic material in the pathogens infecting both subjects. If subject 2’s pathogens are all from a single branch in the phylogeny of subject 1, that can be compelling evidence for transmission from 1 to 2. On the other hand, if subject 2’s pathogens form a clade that is outside that of subject 1, the evidence points to subject 2 being infected by someone else (Box Figure 8.4.1).

Michael Metzker, a geneticist at Baylor College of Medicine, and his colleagues investigated the Padieu case with this method (Scaduto et al. 2010). They sequenced genetic material from Padieu’s HIV infection, along with

Box Figure 8.4.1 Molecular phylogenies provide insights that can allow scientists to reconstruct how pathogens spread from one person to another. This evidence is now being used in criminal trials. (Data from Bhattacharya 2014)
viruses from the six infected women. The result was the phylogeny shown in Box Figure 8.4.2. The HIV from all six women formed branches nested within Padieu’s clade. Each woman was infected by a different virus, in other words, but all the viruses evolved within Padieu’s body.

Padieu was convicted and sentenced to 45 years in prison. The case is one of a growing number of examples of “forensic phylogenetics” making its way into the courtroom (Bhattacharya 2014).

Box Figure 8.4.2  HIV viruses sampled from Philippe Padieu are marked CC01. All of the samples from Padieu’s HIV-infected partners (CC02–CC07) were nested within the CC01 clade. Such a pattern is consistent with the six women getting infected by Padieu. Because this was used as evidence in an assault trial, the strength of support for the tree topology was critical. As we saw with Figure 8.9, scientists often indicate how strongly their data support particular nodes in a phylogenetic tree by placing the node support results right next to the nodes. Here, the numbers depict Bayesian posterior probabilities (1.0 equals 100% probability), and maximum likelihood bootstrap values (% of sampled trees retaining the node). (Data from Scaduto et al. 2010)

Substitutions
One way that genetic mutations can escape the action of selection is by not affecting an organism’s phenotype (see Box 6.4). As we saw in Chapter 5, much of the variation in their DNA is hardly affected by selection at all (Nei 2005).
A synonymous substitution is a substitution that does not alter the amino acid sequence of a protein. Because synonymous substitutions do not affect the protein an organism produces, they are less prone to selection and often free from selection completely.

A nonsynonymous substitution is a substitution that alters the amino acid sequence of a protein. Nonsynonymous substitutions can affect the phenotype and are therefore more subject to selection.

Noncoding DNA—including most pseudogenes—has no known function. As far as we know, these regions of DNA are simply transmitted as baggage from one generation to the next. Mutations to nonfunctional swaths of sequence are not likely to affect the phenotypes of the individuals that carry them, and as such, they are not likely to be exposed to selection.

Even mutations to protein-coding genes can sometimes escape the action of selection. Mutations to a protein-coding gene may fail to change a protein thanks to the redundancy built into the genetic code. Several different codons may encode the same amino acid (see Figure 5.8). A mutation may switch one codon to another without changing the corresponding amino acid. Scientists call this type of mutation a synonymous (or “silent”) substitution.

As we’ll see, synonymous substitutions are much less subject to selection than nonsynonymous substitutions, which replace one amino acid with another. But that does not mean synonymous substitutions are completely immune to selection’s effects. Synonymous substitutions may affect how efficiently a particular protein is translated, even if it does not alter the resulting structure of the protein itself (Tuller et al. 2010). In the next chapter, we’ll discover how altered levels of expression of genes can have important effects on phenotypes and fitness. Finally, a mutation that does change an amino acid in a protein may still fail to change the function of the protein.

In 1968 Motoo Kimura, a biologist at the Japanese National Institute for Genetics, produced the first formal neutral theory of molecular evolution. Although natural selection could produce phenotypic adaptations, Kimura argued, much of the variation in genomes was the result of genetic drift (Kimura 1968, 1983). From this mathematical theory, Kimura and other researchers constructed hypotheses they could test. They predicted, for example, that neutral mutations would become fixed in populations at a roughly regular rate. When a population split into two lineages, each lineage would acquire its own unique set of neutral mutations. The more time that passed after the lineages diverged, the more different mutations would be fixed in each one.

In the 1970s, Walter Fitch of the University of Wisconsin and Charles Langley of the National Institute of Environmental Health Sciences in North Carolina found some compelling evidence for neutral evolution by comparing proteins from 17 mammals (Langley and Fitch 1974). They examined one protein in particular, known as cytochrome c, and mapped its genetic sequence in humans, horses, and other species. From these results, they determined how many mutations had arisen in each lineage. Fitch and Langley then asked paleontologists to estimate when those lineages had split based on the fossil record, and they drew a graph to compare the two sets of results.

As Figure 8.13 shows, they discovered that the more distantly related two species were, the more mutations had accumulated in each lineage since they split from a common ancestor. The graph was especially striking because the relationship was so linear. Mutations became fixed in the lineages with almost clocklike regularity.

In recent years, scientists have been able to carry out large-scale surveys of the rate of nucleotide substitutions in different classes of DNA sequences. As shown in Figure 8.14, different types of sites experience different rates of molecular evolution. Pseudogenes evolve far faster than nonsynonymous sites in protein-coding genes, for example. These differences also support the neutral theory of evolution because most pseudogenes no longer encode proteins or RNA molecules that are important to the fitness of an organism.

The precise relationship between natural selection and neutral evolution is a complex one that scientists are still exploring. But it is clear that neutral evolution has played a major role in how genomes got to be the way they are today. Neutral theory also provides evolutionary biologists with powerful tools for investigating other aspects of evolution. In the following sections, we will consider two of these applications: a “molecular clock” for determining the age of branching events and a null hypothesis for recognizing cases of natural selection.
Because base pair substitutions accumulate at a roughly clocklike rate, scientists can use mutations to tell time. By counting the number of base pair substitutions in a species’ cytochrome c gene, for example, it’s possible to estimate how long ago its ancestors branched off from our own. Scientists refer to this method of tracking time as the molecular clock.

To use molecular clocks, scientists must first calibrate them. If they can date an event relevant to the history of a group, such as the origin of an island or the age of a fossil, then they can calculate the rate at which nucleotides are being substituted over time for the genes and lineages they are studying.

Scientists must also select the most appropriate genetic material for a given molecular clock study (Moorjani et al. 2016). As Figure 8.14 illustrates, different segments of DNA evolve at different rates. To measure the divergence of species separated by hundreds of millions of years, a slow-evolving segment of DNA will provide greater accuracy than a fast-evolving one because it’s likely to have accumulated less noise due to homoplasy (Section 8.4). Faster evolving regions, on the other hand, permit scientists to date events that unfolded over much shorter time scales—in some cases, mere decades.

As we saw earlier, scientists have used molecular phylogenetics to trace the origin of HIV to viruses that infect apes in central Africa. Once the researchers had a robust phylogeny, they could begin using it to estimate when the viruses shifted to human hosts. In 2000, researchers based at Los Alamos National Laboratory compared the

---

**The Molecular Clock**

A molecular clock is a method used to determine time based on base pair substitutions. Molecular clocks use the rates of molecular change to deduce the divergence time between lineages in a phylogeny, for example. They work best when they can be “calibrated” with other markers of time, such as fossils with known ages and placements.
RNA from 159 HIV-1 viruses that had been isolated from blood samples collected from patients at different times during the previous two decades (Korber et al. 2000). They calculated the most likely rate at which the virus genes mutated, based on how much the viruses differed from one another and how old they were.

Rather than requiring that the viruses all follow one clock strictly, the scientists allowed the mutation rate to vary from branch to branch, and even from site to site within the genes (this method is called a relaxed molecular clock). They estimated from the isolates collected during the 1980s and 1990s that the common ancestor of HIV-1 existed sometime between 1915 and 1941.

Researchers later determined a more precise estimate of the origin of HIV, thanks to the discovery of historical samples of HIV. To understand the evolution of the disease, Michael Worobey of the University of Arizona traveled to Africa, where the animal hosts of HIV’s ancestors live. Worobey wondered if HIV might have been infecting people in Africa long before the disease was identified in the 1980s. While visiting hospitals in Kinshasa, the capital of the Democratic Republic of the Congo, he discovered blood samples dating back decades (they were preserved in paraffin). In one of those samples, he found genetic material from HIV (Worobey et al. 2008).

As expected, the early generations of the virus had acquired fewer mutations than more recent ones. By comparing the viruses, Worobey and his colleagues were able to determine a more precise estimate for the origin of HIV. Using a molecular clock approach, they showed that the disease likely emerged early in the twentieth century (Figure 8.15).

These molecular clock studies, combined with the phylogeny described earlier in this chapter, help us understand how HIV evolved. Hunters in West Africa have a long tradition of killing primates to eat or to sell in village markets. The hunters occasionally would have been exposed to various SIV strains. Some of the viruses that infected them might have replicated slowly, but they soon died out.

In the early 1900s, however, things changed. French and Belgian colonists established railways and extracted timber and other resources from deep within African forests. People had more contact with each other in the region, and the growing population drove an increased demand for bushmeat. In this new environment, SIV crossed over into humans and established itself as a new, devastating human pathogen.
Neutral mutations accumulate in a clocklike fashion in genomes. Scientists can use molecular clocks to estimate the origin of diseases and major clades.

### 8.7 Footprints of Selection

Natural selection leaves behind traces in genomes, like footprints in wet sand. Scientists can use a number of different methods to throw a light on those footprints, long after these episodes of selection have passed. Table 8.1 summarizes these methods. We’ll take a closer look in this section at a few of these methods and some examples of the insights they can provide.

#### Table 8.1 General Approaches and Timing of Detecting Selection in Genome-Wide Selection Studies

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Signatures</th>
<th>Scope of the Comparison</th>
<th>Selection Detected</th>
<th>Time Frame (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative</td>
<td>Reduction in the interspecific sequence divergence around a selected region relative to divergence of homologous regions genome-wide (Mayor et al. 2000; Ovcharenko et al. 2004) or when compared with a third species (Tajima 1993)</td>
<td>Between species</td>
<td>Positive, purifying</td>
<td>Greater than 1 million</td>
</tr>
<tr>
<td>Increased function-altering substitution rates</td>
<td>Elevated ratio of nonsynonymous (dN) to synonymous (dS) changes (dN/dS) in coding regions of selected genes (Nielsen and Yang 1998; Yang and Nielsen 1998)</td>
<td>Within a species</td>
<td>Positive</td>
<td>Greater than 1 million</td>
</tr>
<tr>
<td>Interspecies divergence versus intraspecies polymorphism</td>
<td>Reduction in the ratio of intraspecific diversity to interspecific divergence (Hudson et al. 1987; McDonald and Kreitman 1991)</td>
<td>Between species</td>
<td>Positive</td>
<td>Greater than 1 million</td>
</tr>
<tr>
<td>Population based</td>
<td>A significant decrease in genetic variation (often measured as heterozygosity) around the selected site relative to its chromosomal neighborhood or genome-wide (Oleksyk et al. 2008)</td>
<td>Within a population</td>
<td>Positive</td>
<td>Less than 200,000</td>
</tr>
<tr>
<td>Differentiating between populations ($F_{ST}$)</td>
<td>An increase or decrease in population differentiation in genomic regions under selection relative to the rest of the genome (Akey et al. 2002; Beaumont and Balding 2004)</td>
<td>Between populations</td>
<td>Positive, balancing</td>
<td>Less than 80,000</td>
</tr>
<tr>
<td>Extended linkage disequilibrium segments</td>
<td>Extended linkage disequilibrium producing remarkably long haplotypes around the beneficial single-nucleotide phase polymorphism (Tishkoff et al. 2001; Sabeti et al. 2002; Voight et al. 2006)</td>
<td>Within a population</td>
<td>Positive</td>
<td>Less than 30,000</td>
</tr>
</tbody>
</table>

(Data from Oleksyk et al. 2010)
Linkage Disequilibrium

When a neutral mutation arises, especially in a large population, it may take a very long time for it to reach a high frequency through drift. By the time these alleles wander to fixation, recombination has had plenty of opportunity to mix and match the surrounding stretches of the chromosome, eroding any physical linkage among the new mutation and alleles at adjacent loci. Only the actual mutation becomes fixed, having reached linkage equilibrium.

When an allele experiences strong natural selection, on the other hand, it can spread quickly through a population. When this sweep occurs faster than recombination can separate the allele from the nearby regions of the genome, it leaves a footprint of natural selection known as a selective sweep. Alleles that happened to be sitting on the same chromosome when the mutation occurred get pulled along for the ride—as the new mutation becomes more common, so too do these other alleles—in a phenomenon known as genetic hitchhiking. Strongly selected alleles frequently will be found in a population surrounded by the same set of alleles at neighboring locations (Figure 8.16).

Selective sweep describes the situation in which strong selection can “sweep” a favorable allele to fixation within a population so fast that there is little opportunity for recombination. In the absence of recombination, alleles in large stretches of DNA flanking the favorable allele will also reach high frequency.

Genetic hitchhiking occurs when an allele increases in frequency because it is physically linked to a positively selected allele at a nearby locus.

**Figure 8.16** Scientists can detect the signature of natural selection in an allele by comparing its neighboring alleles in different individuals. Strong natural selection favoring one allele will spread its entire neighborhood to high frequencies in a population before recombination can separate the adjacent loci. A: Each line represents a segment of DNA of one individual in the population. Circles represent nucleotide bases unique to that individual. A new mutation (red star) arises in one individual and raises its fitness. B: The same population, a number of generations later. Individuals who inherited a segment of DNA with the new mutation had higher fitness. The mutation increased in frequency, carrying along its neighboring DNA. As a result, this particular recombinant will be unusually abundant in the population. C: Scans across the genome examining patterns of genetic variation will detect such a selective sweep as a “valley” in the amount of standing genetic variation. Loci adjacent to the selected allele will be less variable than expected when compared with samples of loci from other parts of the genome.
Linkage disequilibrium has revealed a striking case of natural selection in humans: the ability of some people to digest milk as adults (Figure 8.17).

Humans are mammals, and one of the hallmarks of living mammals is the production of milk. Milk is rich in a sugar called lactose, and young mammals produce an enzyme called lactase to break it down into simpler sugars they can digest. Around the time young mammals are weaned, they typically stop producing lactase in their guts because they stop drinking milk. Natural selection should favor this shift because it means that mammals don’t waste energy making an enzyme with no advantage.

About 70% of humans also stop producing lactase in their intestinal cells during childhood. As a result, they can digest milk when they’re young, but they have a difficult time with it when they’re adults. Lactase builds up in their guts, spurring the rapid growth of bacteria that feed on the sugar. The waste released by the bacteria causes indigestion and gas. In about 30% of people, however, cells in the gut continue to produce lactase into adulthood. These people can consume milk and other dairy products without any discomfort because they can break down the lactose, leaving less of the sugar for the gas-generating bacteria to feed on. The difference between lactose-tolerant and lactose-intolerant people is largely due to alleles of the lactase gene, LCT (Swallow 2003).

To understand how 30% of people ended up with alleles for lactose tolerance, we must take a look at the history of cattle. Starting about 10,000 years ago, humans began to domesticate cattle in northwest Europe, East Africa, and certain other regions, leading to a dramatic change in their diet. Now energy-rich milk and milk-based foods were available well into adulthood.

The geography of lactose tolerance matches the geography of domestication fairly well (Figure 8.18). An LCT allele for lactose tolerance (called LCT*P) is most common today in northwest Europe—where cattle were domesticated 10,000 years ago—and rarest in southeast Europe, the farthest point in Europe from that origin. Scientists have also compared the frequency of the allele in traditional milk-drinking societies and non-milk-drinking ones in the same countries. LCT*P is generally much more common in the milk drinkers (Swallow 2003). If LCT*P had spread due merely to genetic drift, we would not expect such a strong association between the presence of both cattle herding and the allele. Instead, this pattern points strongly to natural selection.
Sara Tishkoff and her colleagues have found another line of evidence in favor of natural selection (versus random genetic drift) by comparing the DNA of individuals in milk-drinking societies and searching for a selective sweep. Specifically, she looked for this signal of natural selection around the LCT gene in two milk-drinking populations: East Africans and Europeans. The results are shown in Figure 8.19. Strong natural selection has preserved large segments of homologous DNA around LCT. However, a different allele was favored by natural selection in each population of humans. In other words, a mutation arose independently in each population that conferred lactose tolerance and then spread rapidly in both continents (Tishkoff et al. 2007).

We can combine this evidence to come up with a hypothesis for the origin of lactose tolerance. Originally, we humans had an LCT allele that stopped producing lactase when we outgrew nursing. Sometimes mutations gave rise to LCT alleles conferring lactose tolerance in adults, but they did not raise fitness because feeding on milk as adults was rare. In cattle-herding cultures, however, milk was plentiful, and the ability to digest milk brought huge benefits. People who could get protein and other nutrients from milk were more likely to survive and to pass on their mutant copy of LCT to their offspring. We will return to this LCT story in Chapter 18, to see how ancient DNA in fossils is enriching our understanding of its evolution.

**F** _ST Outlier Methods

We saw in Chapter 6 that most natural populations are subdivided into populations dispersed across the landscape. The frequencies of alleles in these populations are the product of opposing forces. Gene flow among the populations works to homogenize their allele frequencies. Meanwhile, drift and selection act within particular populations, causing allele frequencies to diverge from one population to the next.

In Box 6.8, we discussed how population geneticists characterize the extent of subdivision among populations using _F_ _ST_. _F_ _ST_ ranges from 0 (fully homogenized) to 1 (fully segregated). Sequencing a few dozen loci allows scientists to use _F_ _ST_ to calculate a reliable estimate of gene flow.

Initially, scientists used estimates of _F_ _ST_ to measure gene flow between populations. For this research, they sampled a few dozen loci. Now, thanks to the genomics revolution, scientists can estimate _F_ _ST_ for thousands of loci. This wealth of information allows scientists to use _F_ _ST_ to address an entirely new question: how natural selection acts on populations.

When scientists calculate _F_ _ST_ for so many loci, they can determine the overall differentiation between two populations. In some cases, one locus may be an outlier...
African G/C-14010

Eurasian C/T-13910

Figure 8.19 A: Sarah Tishkoff of the University of Pennsylvania and her colleagues examined genetic linkage in Africans to detect natural selection around LCT. They compared 123 people from Kenya and Tanzania with an allele for lactose tolerance and one for intolerance. The people with the lactose tolerance allele (green lines) share much larger segments of homologous DNA around the gene, represented by the length of lines, than those with the alternative lactose-intolerant allele (red lines). B: Working with 101 people of European and Asian descent, Joel Hirschhorn of Harvard and his colleagues also found that an allele for lactose persistence was surrounded by large swaths of homologous DNA (orange lines). Note that the mutations for these two alleles are located in different parts of the same gene. They are lined up in this graph simply to show the different sizes of the hitchhiking regions. (Data from Tishkoff et al. 2007)

The \( F_{ST} \) outlier method detects loci with allele frequencies that are more different than expected between populations. These outlier loci are likely to be near to regions of the genome experiencing strong selection.

compared to the others. In other words, the frequency of an allele at that locus differs between populations to a greater degree than expected by chance. Natural selection is the only mechanism that can produce such extreme outliers. Searching across the genome for such \( F_{ST} \) outliers can point researchers to regions of the genome under strong contemporary selection.

Scientists used \( F_{ST} \) to discover one such case of intense selection in Tibet. At least 30,000 years ago, humans migrated onto the Tibetan Plateau, a vast uplifted swath of Central Asia often called the “Roof of the World.” Spanning more than 2,500,000 square kilometers and averaging 4500 meters in elevation, the Tibetan Plateau is home to large herds of yak, gazelle, and antelope, as well as locally adapted populations of wolf and pika.

All of these species faced the same challenge when they colonized this high, barren landscape: there wasn’t enough oxygen. The partial pressure of oxygen in air drops
as elevation increases, and this means there is less in every breath available for animals to use. Mountain climbers know this all too well, and this is the reason they acclimate at high-elevation base camps for several weeks before a significant ascent. Over a period of a week or two their bodies acclimate, upregulating synthesis of cofactors that bind to the hemoglobin in their blood, for example, making it more readily accessible to their oxygen-starved tissues. Faster breathing rates can bring in more oxygen, too. But these are just short-term strategies for coping with low oxygen. Natural selection has produced other solutions for long-term survival.

Several teams of researchers set out to identify the genes involved with adaptation to high elevation using the \( F_{ST} \) outlier approach. One team, led by Shuhua Xu at the Chinese Academy of Sciences, conducted a genome-wide screen of 1,000,000 genetic markers (SNPs) sampled from 46 high-elevation Tibetans and compared this with samples from 92 low-elevation Han Chinese. They then marched through the genome, comparing the relative frequencies of alleles at each locus across the high- and low-elevation populations.

When Xu and his team compared allele frequencies for Han Chinese and Tibetans, they identified two strong outliers, located next to the \( EPAS1 \) and \( EGLN1 \) genes, respectively (Figure 8.20; Xu et al. 2010). Both of these genes are known to affect oxygen physiology, and the particular alleles identified in Tibetans have since been shown to confer a performance advantage at high elevation (Moore 2017).

Two twists have recently been added to this story. First, a detailed study of the history of the Tibetan \( EPAS1 \) allele found that regions of the genome surrounding this allele were more similar to the genomes of Denisovans, a hominin sister species to humans that we’ll revisit in Chapter 17, than they were to \( EPAS1 \) alleles from other human populations. The gene tree for the Tibetan \( EPAS1 \) allele did not match our species tree. It now appears that Denisovans first possessed the \( EPAS1 \) allele favored by selection at high elevation. The ancestors of Tibetans must have encountered and
interbred with the Denisovans, and their \( EPAS1 \) allele was introgressed into the human genome as a result (Huerta-Sánchez et al. 2014).

The second twist involves studies of Tibetan mastiff dogs. Guo-Dong Wang, also at the Chinese Academy of Sciences, and colleagues used the \( F_{ST} \) outlier approach to compare allele frequencies in the genomes of domesticated dogs sampled from the Tibetan Plateau and from the Chinese lowlands. Not only did they also discover several candidate genes likely to be associated with local adaptation to high elevation, but one of them, \( EPAS1 \), was the same gene as in humans (Wang et al. 2014). It turns out that Tibetan domesticated dogs also likely acquired their beneficial \( EPAS1 \) allele through introgression—in this case from Tibetan gray wolves (Miao et al. 2016).

**\( dN/dS \)**

Both selective sweeps and \( F_{ST} \) outlier methods are able to detect recent selection because the footprints of selection depend on there still being linkage disequilibrium between the mutation under selection and markers located at nearby loci. But, like tracks on a beach, these signals decay with time. Even strongly selected mutations become uncoupled from alleles at adjacent loci through recombination, causing the footprint of selection to disappear.

Here we will consider other methods, based on molecular phylogeny, that scientists can use to detect natural selection that took place millions of years ago.

These methods also use neutral evolution as their null hypothesis. Scientists start out by assuming that any variations they find in homologous segments of DNA are the result of neutral evolution. If they test that hypothesis and reject it, the result supports the interpretation that selection is responsible. (This approach is similar to the way scientists use the Hardy-Weinberg equilibrium as a null hypothesis to detect evidence of selection in genotype frequencies; see Section 6.3.)

One way to test DNA is to compare the substitutions that occur in nonsynonymous sites to those that occur in synonymous sites. We would expect the difference between these two kinds of substitutions to reflect the balance between selection and neutral evolution because only those occurring in nonsynonymous sites would be consistently subject to natural selection. Let’s consider a neutrally evolving pseudogene. It encodes no useful protein or RNA molecule. As a result, selection cannot act on any mutation that it acquires. We would expect that synonymous and nonsynonymous mutations are equally likely to become fixed through genetic drift. To estimate these probabilities, we can calculate the number of nonsynonymous substitutions per nonsynonymous site in the pseudogene (known as \( dN \)) and the number of synonymous substitutions per synonymous site in the pseudogene (known as \( dS \)). Under neutral evolution, we would expect that \( dN = dS \) (Figure 8.21; Bell 2008).

We would reject this hypothesis if we found that synonymous and nonsynonymous mutations did not occur equally as predicted by neutral evolution. Consider, for instance, a gene that undergoes strong positive selection (Chapter 6). It acquires a replacement (nonsynonymous) mutation that alters the structure of the encoded protein in a way that improves its performance. Because this allele is beneficial, it increases in frequency faster than synonymous alleles, the frequencies of which change only due to chance. As mutations continue to arise, beneficial ones will be pulled to fixation by selection, whereas neutral mutations will not. Positive selection will thus produce a gene in which there are more nonsynonymous mutations than would be expected through neutral evolution (\( dN > dS \)).

Another deviation from the null hypothesis occurs if nonsynonymous mutations are fewer than expected. This purifying selection results when a segment of DNA plays an essential role that is easily disrupted by mutations. A gene may encode a protein, for example, that cannot function if even a single amino acid is altered. Natural selection will eliminate alleles of such a gene with harmful mutations that lower fitness. Synonymous mutations, on the other hand, will remain hidden from selection because they don’t alter the protein. Alleles with synonymous mutations will become more frequent, thanks to genetic drift. In other words, \( dS > dN \).
Evolutionary biologists have found that selection has to be very strong to create a clear signal in the difference between $dS$ and $dN$ (Charlesworth and Charlesworth 2010). Under certain conditions, they can use other tests to detect weaker footprints of selection. One of these is called the McDonald-Kreitman test, or MK test for short (McDonald and Kreitman 1991). To carry out the MK test on a gene in a particular species, evolutionary biologists compare its alleles within that species, and they also compare it to the homologous gene in other species. If the gene has experienced neutral evolution, then the ratio of nonsynonymous to synonymous substitutions across species should be the same as the ratio of nonsynonymous to synonymous polymorphic loci within the species.

If, on the other hand, the gene has experienced positive selection in the recent history of the focal species, then beneficial mutations will have rapidly increased and become fixed. These beneficial mutations are more likely to be nonsynonymous substitutions that improve the structure or function of the corresponding protein. As a result, scientists will find a higher ratio of nonsynonymous substitutions to synonymous ones between the focal species and a closely related species.

Selection will have a different effect on alleles within the focal species. Most of the standing genetic variation within the species will be the result of synonymous substitutions. As we saw earlier, selection will lead to the fixation of beneficial nonsynonymous substitutions. As a result, the ratio of nonsynonymous to synonymous fixed substitutions compared across species will be higher, whereas the same ratio calculated for polymorphic loci within the species will be lower. Such a pattern can be
interpreted as evidence for positive selection fixing beneficial mutations between the species.

The BRCA1 gene offers a fascinating case study in natural selection revealed by nonsynonymous mutations. As we discussed earlier, BRCA1 is best known as a gene associated with breast cancer. But when it isn’t causing cancer, it serves a number of vital functions. It oversees repairs to damaged DNA, and it helps control the complex molecular events that allow one cell to divide into two.

To explore the evolutionary history of this medically important gene, University of Texas biologist Sara Sawyer and her colleagues studied the synonymous and nonsynonymous mutations in BRCA1 (Lou et al. 2014). They compared orthologs in 23 species of primates.

On many branches, the researchers estimated that the $dN/dS$ ratio was well below 1. This could be the result of negative selection eliminating nonsynonymous mutations that disrupted BRCA1’s function. But on a few branches, the $dN/dS$ value was well above 1, indicating positive selection. Intriguingly, one branch had the highest value of all. Figure 18.22 shows the results of Sawyer’s analysis for humans and our closest ape relatives.

Sawyer and her colleagues found that we have acquired 22 nonsynonymous substitutions since our ancestors split from the ancestors of chimpanzees and bonobos more than 7 million years ago. By contrast, humans have only three fixed synonymous mutations in the BRCA1 gene. What makes this discovery all the more intriguing is that many of these changes to BRCA1 result in a significantly elevated risk for breast cancer.

Sawyer and her colleagues speculate that viruses are the solution to this paradox. When cells divide and make new copies of their DNA, some viruses can slip their own genetic material into our cellular machinery to create new copies of themselves. Mutations to BRCA1 may allow the gene to shut viruses out of this process. But viruses may then evolve new adaptations to evade the BRCA1 protein. As we’ll see in Chapter 15, this open-ended back-and-forth form of evolution is quite common in parasites and their hosts. The benefits of this positive selection may be so great that they outweigh the increased risk of cancer, much like the trade-off between malaria resistance and sickle-cell anemia that we explored in Chapter 6.

---

**Figure 8.22** Certain mutations to the BRCA1 gene create a significant risk for hereditary breast cancer. Surprisingly, the gene has undergone strong positive selection in our own lineage, as well as the lineage giving rise to chimpanzees and bonobos. The number on the top of each branch is the $dN/dS$ ratio. On the bottom of each branch are the number of nonsynonymous and synonymous substitutions estimated to have been fixed on each lineage. (Data from Lou et al. 2014)
• The neutral theory of molecular evolution describes the pattern of nucleotide sequence evolution under the forces of mutation and random genetic drift in the absence of selection.
• The neutral theory predicts that neutral mutations will yield nucleotide substitutions in a population at a rate equivalent to the rate of mutation, regardless of the size of the population.
• As long as mutation rates remain fairly constant through time, neutral variation should accumulate at a steady rate, generating a molecular signature that can be used to date events in the distant past.
• Positive selection and purifying selection both leave distinctive signatures in nucleotide or amino acid sequences that can be detected using statistical tests.

8.8 Genome Evolution

Some evolutionary biologists examine the evolution of individual genes; other researchers look at the genome as a whole. As we saw in Chapter 5, genomes can vary tremendously in size. Many evolutionary factors are at play in determining the genome size in each species. As Figure 8.23 shows, the size of bacterial genomes is proportional to the number of genes in each species. That’s because bacterial genomes are mostly composed of genes.

Bacteria increase their genomes by gaining new genes. They have several mechanisms for doing so. An accidental duplication of a segment of DNA can create an extra copy of a gene. As we’ll discuss in more detail in the next chapter, duplicated genes can diverge, acquiring new functions. Bacteria also acquire new genes from other bacteria through horizontal gene transfer.

Bacterial genomes can shrink as well. Accidental deletions of DNA may eliminate genes; if the loss of those genes doesn’t reduce the fitness of bacteria, they may disappear entirely from a population. Scientists have found that certain kinds of niches favor bacterial genomes of different sizes. Free-living bacteria typically have relatively large genome lengths, indicating that there is more pressure to reduce genome size. On the other hand, have large amounts of noncoding DNA. Eukaryote genomes can become very large without accumulating a proportionately large number of genes. (Data from John McCutcheon)
Genomes, for example, with a large repertoire of genes to cope with a variety of conditions. Some bacteria have become restricted to living in hosts, either as pathogens or as symbionts. With a reliable supply of amino acids and other nutrients from their hosts, many genes become less essential. Mutations that create pseudogenes are less likely to lower fitness. Subsequently, these noncoding regions may be deleted altogether, shrinking the genome (Figure 8.24). As we’ll see in Chapter 15, the genomes of bacteria that exist inside hosts for many millions of years can shrink drastically to merge with the cells they inhabit.

Eukaryotes, on the other hand, don’t have genomes that correlate tightly in size with their number of genes. Genomes with roughly the same number of genes can have vastly different sizes. Deletions can shrink the size of eukaryote genomes, but many processes can increase them. Like prokaryotes, eukaryotes can experience gene duplication. Even their entire genome can be duplicated.

Eukaryotic genomes also contain large amounts of non-protein-coding DNA. Mobile elements (Chapter 5), for example, proliferate in a parasitic fashion. Some researchers think that the dangers of these mobile elements are so great that eukaryotes have evolved defense systems to prevent mobile elements from replicating, thus protecting vital parts of the genome from the insertion of new copies of mobile elements (Slotkin and Martienssen 2007).

Why the genome structure of eukaryotes is so different from that of bacteria is not yet clear. Michael Lynch of Indiana University has proposed that the key factors are that the cells of eukaryotes were much larger than those of bacteria, and that the populations of eukaryotes were smaller (Lynch 2007). Genetic drift in those small populations could have allowed the first mobile elements to invade their genomes, despite their deleterious effects. Ever since that happened, mobile elements and their hosts have been locked in an evolutionary battle.

Figure 8.24 Bacteria can experience drastic reduction in genome size as they evolve into symbionts that live inside hosts. Deletions remove the many genes that are no longer essential. This reduction can continue for tens of millions of years. (Data from McCutcheon and Moran 2012)
The variation in genome sizes of eukaryotes is striking and puzzling. Why, for example, do genome sizes in animals vary 6650-fold (Gregory 2014)? Salamanders have some of the largest animal genomes, which can be up to 40 times the size of the human genome (Sun et al. 2012). Researchers have offered a number of hypotheses to explain these differences in eukaryotes. Some have suggested that animals with larger cells have larger genomes, for example. But no single hypothesis has gained strong support. As research in genomics advances, scientists may be able to untangle answers to these important questions.

**Key Concepts**

- Bacteria typically have relatively small genomes made up mostly of genes, whereas eukaryotes have genomes that vary greatly in size.
- As more and more genomes are sequenced, our understanding of genome evolution is changing rapidly. Scientists are discovering ways to answer important questions about the role that genome size and architecture play in the origins of our species as well as in other eukaryotes.

**TO SUM UP . . .**

- The relationships among species and other lineages can be inferred from different lines of evidence, such as their DNA, as well as their morphology. These different lines of evidence may yield different phylogenetic hypotheses.
- Scientists use coalescence to trace the alleles of a gene shared by all individuals of a population back to a single common ancestor.
- The phylogeny of a single segment of DNA may be different from the phylogeny of the species that carry it. As a result, scientists use several genes when they examine the phylogenetic relationships among species.
- Phylogenies represent hypotheses describing historical relationships based on currently available evidence. Statistical methods such as maximum parsimony, bootstrapping, distance matrixes, maximum likelihood, and Bayesian approaches help scientists resolve their data and develop phylogenetic hypotheses that can be tested with additional evidence.
- Scientists have used molecular phylogenetics to test hypotheses about evolution, from the origin of tetrapods to the origin of pathogens such as HIV.
- Nucleotide sequences evolve “neutrally” in the absence of selection as a result of mutation and random genetic drift. Much of the human genome shows patterns of neutral evolution.
- Molecular “clocks” allow scientists to estimate the age of common ancestors by comparing the mutations in the same genomic context across a group of organisms.
- Neutral evolution can serve as a null model in tests designed to detect natural selection because it predicts that substitutions that occur in nonsynonymous (replacement) sites and synonymous sites are equally likely to become fixed. If scientists find that the substitutions are not equal, they can reject the null hypothesis and infer that selection is responsible for the differences.
- Statistical tests allow scientists to detect the distinctive signatures in nucleotide or amino acid sequences that result from positive and purifying selection.
- Although scientists have hypothesized that genome size may be related to cell size, understanding the diversity of genome sizes and its role in evolution are fertile areas of study in evolutionary biology.
MULTIPLE CHOICE QUESTIONS  Answers can be found at the end of the book.

1. Why is understanding coalescence important when developing molecular phylogenies?  
   a. Because scientists can’t know the true genealogy of a lineage without coalescing phylogenies to determine which is the most parsimonious.  
   b. Because scientists can sample only a limited portion of the history of any allele.  
   c. Because alleles that change over time are not valuable to developing phylogenies.  
   d. Because scientists can’t possibly determine the genealogy of a lineage from the limited samples available to them.

2. Why don’t all gene trees reflect the phylogeny of species?  
   a. Because the branch lengths of a species tree are usually much longer on average than the coalescence times of the genes being analyzed.  
   b. Because coalescence of specific genes can occur before speciation events.  
   c. Because speciation events can sometimes be very rapid.  
   d. Both b and c.

3. Why might scientists use a statistical tool, such as Bayesian or maximum likelihood analyses, when reconstructing phylogenies?  
   a. Because otherwise scientists can easily misinterpret the outcome.  
   b. Because scientists can specify the parameters of a statistical model and test the capacity of the tool to produce comparable trees.  
   c. Because molecular data can provide both true and false signals of the branching history, and statistical tools can reveal important patterns in the changes that occurred.  
   d. Both b and c.

4. Molecular phylogenies indicate which of the following about HIV?  
   a. The same mutation evolved in three separate lineages of HIV; in each instance, the mutation improved the ability of the virus to infect humans.  
   b. HIV came from a monkey virus that was introduced into people by contaminated vaccinations.  
   c. HIV is a monophyletic strain of lentivirus that infects both humans and chimpanzees.  
   d. The common ancestor of simian immunodeficiency virus and human immunodeficiency virus came from horses.

5. The theory of neutral evolution describes  
   a. the rate of mutation at a site that results from purifying selection, regardless of the size of the population.  
   b. the rate of fixation of alleles at a site in the absence of selection.  
   c. the competition between genetic drift and natural selection within the genome.  
   d. Both a and b.

6. Which of these is a true statement about molecular clocks?  
   a. Molecular clocks use neutral theory to date events within a phylogeny.  
   b. Molecular clocks can be calibrated using fossils of known age.  
   c. Molecular clocks can be affected by the segments of DNA being examined and relative sizes of the populations.  
   d. All of the above.

7. When \( dN > dS \),  
   a. scientists would reject the null hypothesis of neutral evolution because the number of replacement substitutions is greater than expected.  
   b. scientists would accept the hypothesis that the population is undergoing purifying selection because more replacement mutations were found than expected.  
   c. scientists would reject the hypothesis that natural selection took place millions of years ago and is no longer relevant.  
   d. scientists would accept the hypothesis that neutral evolution took place millions of years ago.

8. Which of the following is not true of coalescence?  
   a. The timing of coalescence can depend on whether or not alleles are under selection.  
   b. Positive selection can accelerate the rise in frequency in an allele, leading to a short coalescence.  
   c. Two alleles that experience little selection may coexist longer, and thus the farther back coalescence occurs.  
   d. In some instances, it is impossible to trace the genealogies of two homologous alleles in a population to a common ancestral allele, even if it is possible to trace their histories indefinitely.

9. Why do scientists use several genes when they examine the phylogenetic relationships among species?  
   a. Synonymous substitutions are more likely to be present in multiple genes.  
   b. The phylogeny of a single segment of DNA may be different from the phylogeny of the species that carry it.  
   c. Purifying selection can remove deleterious alleles from a population, and a single segment of DNA may be missing from one gene.  
   d. Scientists are aiming to increase their chances of finding microsatellites, which can be valuable genetic characters for comparing populations.

10. Which of the following is not true of the neutral theory of molecular evolution?  
    a. When neutral variation accumulates at a steady rate, the molecular signature generated is an unreliable measure to date events in the distant past.  
    b. The neutral theory of molecular evolution describes the pattern of nucleotide sequence evolution under the forces of mutation and random genetic drift in the absence of selection.  
    c. The neutral theory predicts that neutral mutations will yield nucleotide substitutions in a population at a rate equivalent to the rate of mutation, regardless of the size of the population.  
    d. As long as mutation rates remain constant, neutral variation is expected to accumulate at a steady rate.
INTERPRET THE DATA  Answer can be found at the end of the book.

11. The figure below shows the estimated times of coalescence for alleles of 24 human autosomal genes and the mitochondrial genome. Most of the genes appear to coalesce at roughly the same time, between 0.5 and 1.5 million years ago. But the coalescent time of MX1 is much deeper. Why might alleles of this gene be so much older than alleles of the other genes?

Time to most recent common ancestor (millions of years)

- Y-DNA
- mt-DNA
- MAO
- FOX
- MSLN/ALAS2
- Xq13.3
- G6PD
- HSS71R2
- APLX
- MC1R
- ECP
- EDN
- AMLEX
- MS205
- HFE
- TNFSF5
- Hbβ
- CYP1A2
- RRM2B
- PDHA1
- FUT6
- LACTASE
- CCR5
- FUT2
- MX1

H. sapiens

a. Because human populations likely went through a bottleneck event when they dispersed out of Africa.
b. Because one or more of the alleles of this gene were introgressed into human populations from a divergent lineage, such as the Neanderthals or Denisovans.
c. Because alleles of this gene are experiencing positive selection.
d. Because human populations have undergone a dramatic recent explosion in numbers.
SHORT ANSWER QUESTIONS  Answers can be found at the end of the book.

12. Why are different statistical methodologies important for developing phylogenies?

13. How is the theory of neutral evolution different from the theory of evolution by natural selection? How is it the same?

14. Why might humans and dogs have used the same genes when they locally adapted to high elevation populations?

15. How did Sarah Tishkoff and her colleagues test the hypothesis that all modern humans are derived from recent African ancestry?

16. Why are conserved genes important when developing phylogenies?

17. Discuss the differences between a gene tree and a species tree.

ADDITIONAL READING


PRIMAR LITERATURE CITED IN CHAPTER 8


Sun, C., D. B. Shepard, R. A. Chong, J. L. Arriaza, K. Hall, et al. 2012. LTR Retrotransposons Contribute to Genomic Gigan-
tism in Plethodontid Salamanders. Genome Biology and Evolu-
tion 4:168–83.
Associates.
Takezaki, N., and H. Nishihara. 2017. Support for Lungfish as the Closest Relative of Tetrapods by Using Slowly Evolving
Ray-Finned Fish as the Outgr
da.
Aposematic Species: A Phylogenetic Analysis of Color Variation
Xu, S., S. Li, Y. Yang, J. Tan, H. Lou, et al. 2010. A Genome-
Wide Search for Signals of High-Altitude Adaptation in
Yang, Z., and R. Nielsen. 1998. Synonymous and Non-synonymous
Rate Variation in Nuclear Genes of Mammals. Journal of
Molecular Evolution 46:409–18.
Using DNA Sequences: A Markov Chain Monte Carlo