7.3 Genetic Drift and Molecular Evolution

The study of molecular evolution began in the mid-1960s, when biochemists succeeded in determining the amino acid sequences of hemoglobin, cytochrome *c*, and other abundant and well-studied proteins found in humans and other vertebrates. These sequences provided the first opportunity for evolutionary biologists to compare the amount and rate of molecular change among species.

Early workers in the field made several striking observations. Foremost among them were calculations by Motoo Kimura (1968). Kimura took the number of sequence differences in the well-studied proteins of humans versus horses and converted them to rates of sequence change over time using divergence dates estimated from the fossil record. He then extrapolated these rates to all of the protein-coding loci in the genome. The result implied that as the two lineages diverged from their common ancestor, mutations leading to amino acid replacements had, on average, risen to fixation once every two years. Given that most mutations are thought to be deleterious, this rate seemed too high to be due to natural selection. Beneficial mutations fixed by natural selection should be rare.

A second observation, by Emil Zuckerkandl and Linus Pauling (1965), was that the rate of amino acid sequence change in certain proteins appeared to have been constant over time, or clocklike, during the diversification of vertebrates. This too seemed inconsistent with natural selection, which should be episodic and correlated with environmental change rather than with time.

Early analyses of molecular evolution suggested that rates of change were high and constant through time. These conclusions appeared to be in conflict with what might be expected under natural selection.

In short, early data on molecular evolution did not match expectations derived from the notion that most evolutionary change was due to natural selection. But if natural selection does not explain evolution at the molecular level, then what process is responsible for rapid, clocklike sequence change? Many researchers believe the answer is genetic drift.

The Neutral Theory of Molecular Evolution

Kimura (1968, 1983) formulated the **neutral theory** of molecular evolution to explain the observed patterns of amino acid sequence divergence. To understand the neutral theory's central claim, note that with respect to effect on fitness, there are three kinds of mutations. Some mutations are deleterious, some are neutral, and some are beneficial. Mutations that are deleterious tend to be eliminated by natural selection and thus contribute little to molecular evolution. Mutations that are neutral (or nearly so—more on that later) rise and fall in frequency as a result of genetic drift. Many are lost, but some become fixed. Mutations that are beneficial are often lost to drift while still at low frequency, but otherwise tend to rise to fixation as a result of natural selection. Kimura's neutral theory holds that effectively neutral mutations that rise to fixation by drift vastly outnumber beneficial mutations that rise to fixation by natural selection. Genetic drift, not natural selection, is thus the mechanism responsible for most molecular evolution.

The neutral theory models the fate of new alleles that were created by mutation and whose frequencies change by genetic drift. It claims to explain most evolutionary change at the level of nucleotide sequences.

Based on his view that drift dominates sequence evolution, and on the calculation detailed in Computing Consequences 7.5, Kimura postulated that the rate of molecular evolution is, to a good approximation, equal to the mutation rate.

Kimura's theory was startling to many evolutionary biologists. Given that drift has a larger influence on allele frequencies in small populations than in large ones, the absence of an effect of population size on the rate of evolution was counterintuitive. So was the assertion that sequence evolution by natural selection was so rare, compared to evolution by drift, as to be insignificant.

Although Kimura's theory appeared to explain why the amino acid sequences of hemoglobin, cytochrome *c*, and other proteins change steadily over time, the theory was inspired by limited amounts of data. How did the neutral theory hold up, once large volumes of DNA sequence data became available?

Patterns in DNA Sequence Divergence

During the late 1970s and 1980s, biologists mined growing databases of DNA sequences to analyze the amounts and rates of change in different loci. They began to see patterns that varied by the type of sequence examined. The most basic distinction was between coding versus noncoding sequences. Coding sequences contain instructions for tRNAs, rRNAs, or proteins; noncoding sequences include introns, regions that flank coding regions, regulatory sites, and pseudogenes. What predictions does the neutral theory make about the rate and pattern of change in different types of sequences, and have they been verified or rejected?

Pseudogenes Establish a Canonical Rate of Neutral Evolution

Pseudogenes are functionless stretches of DNA that result from gene duplications (see Chapter 5). Because they do not encode proteins, mutations in pseudogenes should be neutral with respect to fitness. When such mutations achieve fixation in populations, it should happen solely as a result of drift. Pseudogenes are thus considered a paradigm of neutral evolution (Li et al. 1981). As predicted by the neutral theory, the divergence rates recorded in pseudogenes—which should be equal to the neutral mutation rate ν—are among the highest seen for loci in nuclear genomes (Li et al. 1981; Li and Graur 2000). This finding is consistent with the neutral theory's explanation for evolutionary change at the molecular level. It also quantifies the rate of evolution due to drift. For humans versus chimps, this rate is about 2.5×10^{-8} mutations per nucleotide site per generation (Nachman and Crowell 2000). How do rates of change in other types of sequences compare to the standard, or canonical, rate?

The evolution of pseudogenes conforms to the assumptions and predictions of the neutral theory.

Silent Sites Change Faster than Replacement Sites in Most Coding Loci

Recall (from Chapter 5) that bases in DNA are read in three-letter codons, and that the genetic code contains considerable redundancy. In the portion of the code shown in Figure 7.24a, two codons specify phenylalanine, two specify leucine, and four code for serine. As shown in Figure 7.24b, base-pair changes

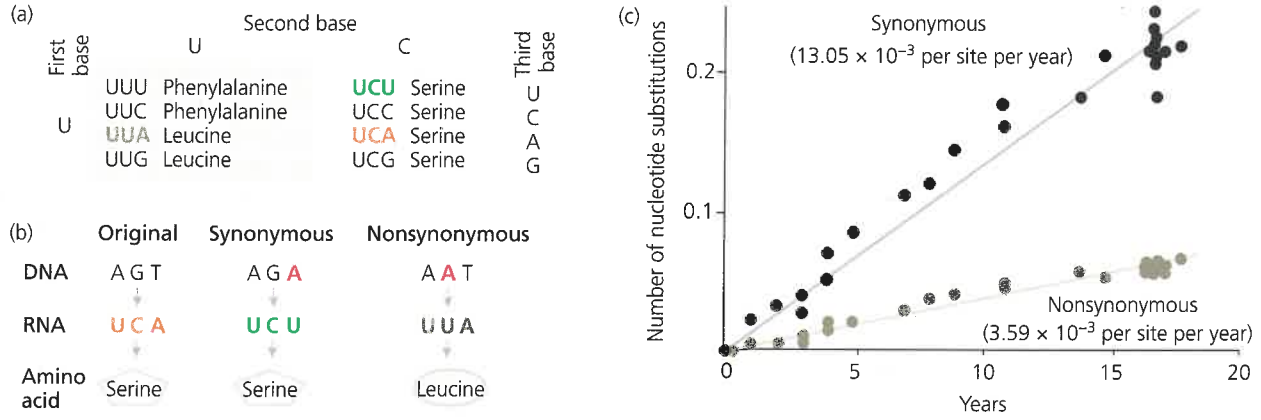

Figure 7.24 Molecular evolution in influenza viruses is consistent with the neutral theory Because the genetic code is redundant (a), there are two kinds of point mutations (b). The neutral theory predicts that both will accumulate by drift, but synonymous substitutions will accumulate faster. (c) Data from the flu virus. From Gojobori et al. (1990).

may or may not lead to amino acid sequence changes. DNA sequence changes that do not result in amino acid changes are called **silent-site** (or **synonymous**) mutations; sequence changes that do result in an amino acid change are called **replacement** (or **nonsynonymous**) mutations.

Figure 7.24c presents data on the rate of silent versus replacement substitution in a gene belonging to the influenza virus, based on comparisons of flu viruses collected over a span of 20 years with a reference sample collected in 1968 (Gojobori et al. 1990). Both kinds of substitution accumulated in a linear, clocklike fashion, but the rate of evolution for silent changes is much higher than the rate of evolution for replacement changes.

This pattern accords with the neutral theory. Silent changes are not exposed to natural selection on protein function, because they do not alter the amino acid sequence. New alleles created by silent mutations should thus increase or decrease in frequency largely as a result of drift. Replacement mutations, in contrast, change the amino acid sequences of proteins. If most of these alterations are deleterious, then most of them should be eliminated by natural selection without ever becoming common enough to be detected. This type of natural selection is called **negative** or **purifying selection,** as opposed to **positive selection** on beneficial mutations. Less frequently, replacement mutations occur that have no effect on protein function and may be fixed by drift.

Natural selection against deleterious mutations is called negative selection.
Natural selection favoring beneficial mutations is called positive selection.

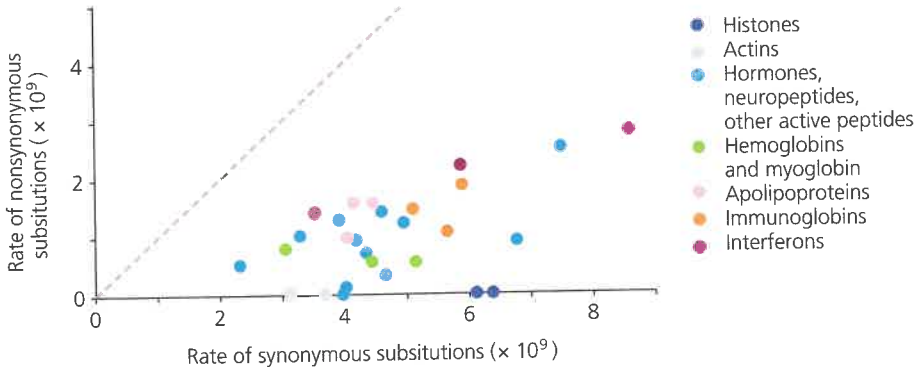

Figure 7.25 Rates of nucleotide substitution vary among genes and among sites within genes Data points report rates of replacement and silent substitutions in protein-coding genes compared between humans and either mice or rats. Units are substitutions per site per billion years. The number of codons compared per gene ranges from 28 to 435. Data from Li and Graur (1991).

Molecular biologists have compared the rate of replacement versus silent substitutions in a great variety of coding loci. In Figure 7.25, the dashed line marks where the data would fall if the nonsynonymous and synonymous substitutions accumulate at equal rates. Genes in which nonsynonymous changes accumulate faster would appear above the line. Genes in which synonymous changes accumulate faster fall below it. In the vast majority of genes studied, the rate of evolution involving silent changes is far higher than the rate of evolution involving replacements.

In most coding sequences, substitution rates are higher at silent sites than at replacement sites. This result is consistent with the notion that molecular evolution is dominated by drift and negative selection.

In a similar vein, Austin Hughes and colleagues (2003) examined the DNA of 102 ethnically diverse humans to quantify the standing genetic diversity at 1,442 single-nucleotide polymorphisms. A single-nucleotide polymorphism is a point in the genome at which some individuals have one nucleotide and other individuals have another. The researchers found lower standing diversity, measured as the fraction of individuals who are heterozygotes, for polymorphisms that involve amino acid changes versus polymorphisms that do not. These results imply that most single-nucleotide mutations that swap one amino acid for another are deleterious and held at low frequency by negative selection.

These observations are consistent with the patterns predicted if most mutations are either deleterious or neutral and drift dominates molecular evolution. They support the central tenet of the neutral theory.

Variation among Loci: Evidence for Functional Constraints

The data in Figure 7.25 contain another important pattern. When homologous coding sequences from humans and rodents are compared, some loci are found to be nearly identical, while others have undergone rapid divergence. This result turns out to be typical. Rates of molecular evolution vary widely among loci.

The key to explaining this pattern is that genes responsible for the most vital cellular functions appear to have the lowest rates of replacement substitutions. Histone proteins, for example, interact with DNA to form structures called nucleosomes. These protein–DNA complexes are a major feature of the chromatin fibers in eukaryotic cells. Changes in the amino acid sequences of histones disrupt the structural integrity of the nucleosome and have ill consequences for DNA transcription and synthesis. In contrast, genes less vital to the cell, and thus under less stringent functional constraints, show more rapid rates of replacement substitutions. When functional constraints are lower, a larger fraction of replacement mutations are neutral with respect to fitness and may fix by drift.

Nearly Neutral Mutations

Although the neutral theory appeared to account for several important patterns in molecular evolution, data indicating clocklike change in proteins compared across species presented a problem. The issue was that the neutral mutation rate v should vary among species as a function of generation time, not clock time. Over a given interval of clock time, more neutral substitutions should accumulate in species with short generation times than in species with long generation times. Contrary to expectation, at least some protein sequence comparisons reveal clocklike change in absolute time—independent of differences in generation time among the species compared. The data points in Figure 7.26 fall along lines, despite comparing humans to species with drastically different generation times.

To account for this observation, Tomoko Ohta and Motoo Kimura (1971; Ohta 1972, 1977) considered how the probability of fixation for a novel mutation depends on the effective population size and the strength of selection. We looked at an example of this relationship in Figure 7.23. If the product of twice the effective population size and the selection coefficient is sufficiently close to zero—because the population is tiny, selection is weak, or both—the probability

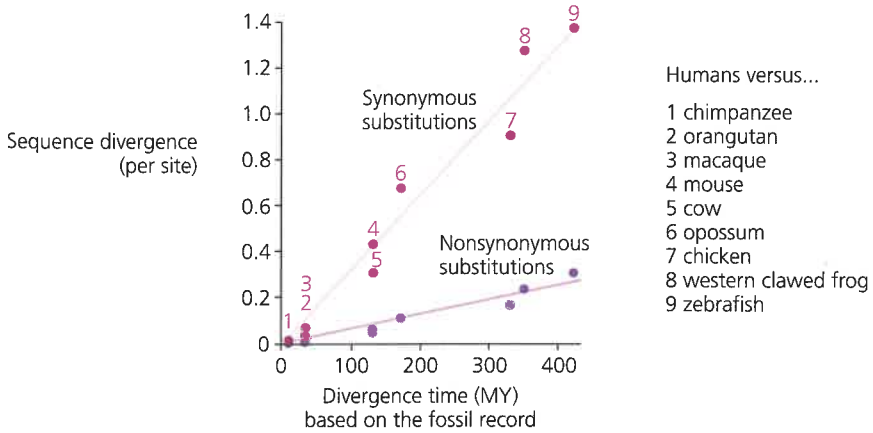

Figure 7.26 The vertebrate molecular clock ticks in calendar time, not generation time The data points, showing sequence divergence versus clock time, fall on lines—regardless of whether they compare humans with other species with long generation times (chimpanzees, orangutans) or short generation times (mouse, zebrafish). Each point represents an average for over 4,000 genes. From Nei et al. (2010).

of fixation is roughly the same as it would be if the mutation had no effect on fitness at all. The allele's frequency will evolve primarily as a result of genetic drift. In population genetics models of evolution in finite populations, neutral alleles and nearly neutral alleles behave the same way.

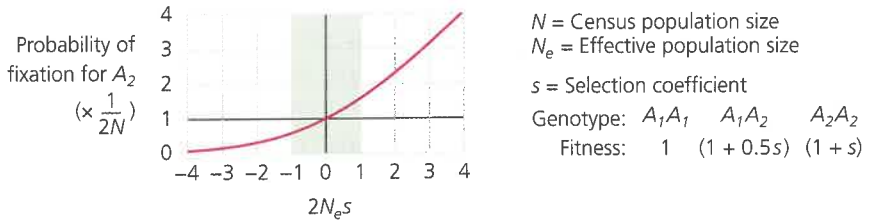

Figure 7.27 A definition of nearly neutral evolution The green band shows a range of values for $2N_es$ over which the frequency of a new mutation changes mostly by genetic drift. After Charlesworth (2009).

We have reproduced part of Figure 7.23 in Figure 7.27. Examination of the figure will reveal that the threshold value of $|2N_es|$ below which we will call a mutation nearly neutral is somewhat arbitrary. It also depends on how the selection coefficient is defined. Ohta and Kimura's (1971) analysis suggests that, with the selection coefficient defined as in the figure, a reasonable criterion is $|2N_es| \leq 1$, or $|s| \leq \frac{1}{2N_e}$. This range is covered by the green band in the graph.

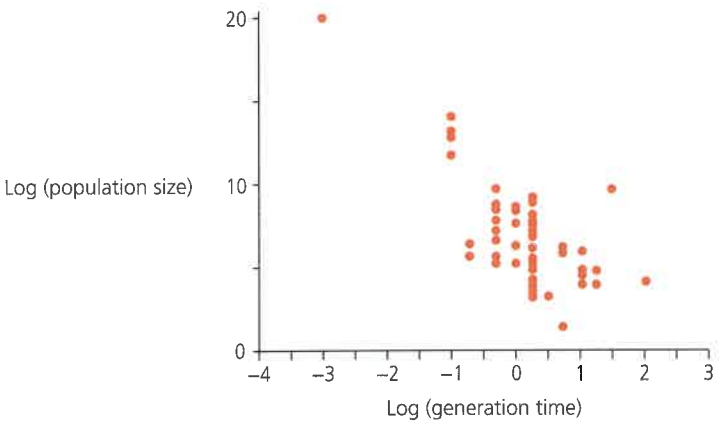

Figure 7.28 Population size versus generation time Across organisms, as generation time goes up, population size goes down. Statistical tests confirm the strong inverse correlation displayed in this log-log plot. From Chao and Carr (1993).

How does the consideration of nearly neutral mutations explain the observation that molecular clocks tick in absolute time rather than in the number of generations? As Lin Chao and David Carr (1993) have shown, there is a strong negative correlation between average population size in a species and its generation time. Species with short generation times tend to have large populations; species with long generation times tend to have small populations (Figure 7.28).

This is important because, Ohta argued, as generation time goes up, population size, and thus $|2N_es|$, go down. As a result, a larger fraction of the mutations that arise—in particular, a larger fraction of the mildly deleterious mutations that are typically abundant in most species—are effectively neutral. Mutations that would be eliminated by purifying selection in a large population of short-lived individuals instead evolve by drift a small population of long-lived individuals. This tends to equalize the rate of evolutionary substitution, measured in absolute time, across species with different generation times.

The nearly neutral model explains why, in some cases, rates of sequence change correlate with absolute time instead of generation time.

Matsatoshi Nei (2005) has suggested that a more biologically meaningful definition of a neutral mutation would consider how much the mean fitness of the

population would change were the mutation to become fixed. If s is defined as in Figure 7.27, Nei would call a mutation effectively neutral if $|s| \leq 0.002$. In population genetics models of extremely large populations, selection this weak can drive an allele to fixation. But the time required for it to do so may be unrealistically long (Nei et al. 2010). Furthermore, for an allele so weakly associated with fitness, the strength and even direction of selection are likely to change over time, across different environments, and on different genetic backgrounds.

The Neutral Theory as a Null Hypothesis: Detecting Natural Selection on DNA Sequences

The neutralist–selectionist controversy is a debate about the relative importance of drift and positive selection in explaining molecular evolution.

Since their inception, the neutral and nearly neutral theories have been controversial (see Berry 1996; Ohta and Kreitman 1996). Discussion has focused on the claims by Kimura (1983) and King and Jukes (1969) that the number of beneficial mutations fixed by positive natural selection is inconsequential compared to the number of mutations that change in frequency under the influence of drift. Is this claim accurate? How can we determine that natural selection has been responsible for changes observed at the molecular level?

When researchers compare homologous DNA sequences among individuals and want to explain the differences they observe, they routinely use the neutral theory as a null hypothesis. The neutral theory specifies the rates and patterns of sequence change that occur in the absence of natural selection. If the changes that are actually observed are significantly different from the predictions made by the neutral theory, and if a researcher can defend the proposition that the sequences in question have functional significance for the organism, then there is convincing evidence that natural selection has caused molecular evolution.

Here we examine a few of the strategies being used to detect molecular evolution due to natural selection. We begin with studies of replacement changes, then explore evidence that many silent-site mutations are also under selection.

Selection on Replacement Mutations

We noted earlier that according to the neutral theory, silent mutations are expected to evolve largely by genetic drift. Replacement mutations are expected either to be deleterious, in which case they are eliminated by negative selection and we will not see them, or to be neutral, in which case they, too, evolve by drift. If the neutral theory is wrong for a particular gene, however, and replacement mutations are advantageous, then they will be rapidly swept to fixation by positive selection. Thus, to find out whether replacements within a particular gene are deleterious, neutral, or advantageous, we can compare two sequences and calculate the rate of nonsynonymous substitutions per site (d_N) and the rate of synonymous substitutions per site (d_S). If we take their ratio, we will get

When sequences evolve by drift and negative selection, synonymous substitutions outnumber replacement substitutions. When sequences evolve by drift and positive selection, replacement substitutions outnumber synonymous substitutions.

$$\frac{d_N}{d_S} < 1 \text{ when replacements are deleterious,}$$

$$\frac{d_N}{d_S} = 1 \text{ when replacements are neutral, and}$$

$$\frac{d_N}{d_S} > 1 \text{ when replacements are advantageous}$$

Austin Hughes and Masatoshi Nei (1988) tested the neutral theory by estimating the ratio of replacement to silent substitutions in genes vital to immune function. When mammalian cells are infected by a bacterium or a virus, they respond

by displaying pieces of bacterial or viral protein on their surfaces. Immune system cells then kill the infected cell, which prevents the bacterium or virus inside the cell from replicating. The membrane proteins that display pathogen proteins are encoded by a cluster of genes called the major histocompatibility complex, or MHC. The part of an MHC protein that binds to the foreign peptide is called the antigen recognition site (ARS). Hughes and Nei (1988) studied sequence changes in the ARS of MHC loci in humans and mice.

When Hughes and Nei compared alleles from the MHC complexes of 12 different humans and counted the number of differences observed in silent versus replacement sites, they found significantly more replacement-site than silent-site changes. The same pattern occurred in the ARS of mouse MHC genes, although the differences were not as great. This pattern could result only if the replacement changes were selectively advantageous. The logic here is that positive selection causes replacement changes to spread through the population much more quickly than neutral alleles can spread by chance.

It is important to note, however, that Hughes and Nei found this pattern only in the ARS. Other exons within the MHC showed more silent than replacement changes, or no difference. At sites other than the ARS, then, they could not rule out the null hypothesis that sequence change is dominated by drift.

In many examples, replacement substitutions outnumber synonymous substitutions—a signature of positive selection.

Research by Gavin Huttley and colleagues (2000) on *BRCA1*, a gene associated with breast cancer, provides another example. *BRCA1* encodes a protein that participates in the repair of damaged DNA (see O'Connell 2010) and in the regulation of programmed cell death during neural development (Pulvers and Huttner 2009). Huttley and colleagues sequenced exon 11 from the *BRCA1* genes of a variety of mammals, then inferred the rates of nonsynonymous and synonymous substitution along the branches of the evolutionary tree that connects the extant species to their common ancestors (Figure 7.29). Along most branches of the phylogeny the value of $\frac{d_N}{d_S}$ was less than one, consistent with the neutral theory. On the branches connecting humans and chimpanzees to their common ancestor, however, $\frac{d_N}{d_S}$ was significantly greater than one. This suggests that the sequence of exon 11 has been under positive selection in the ancestors of today's humans and

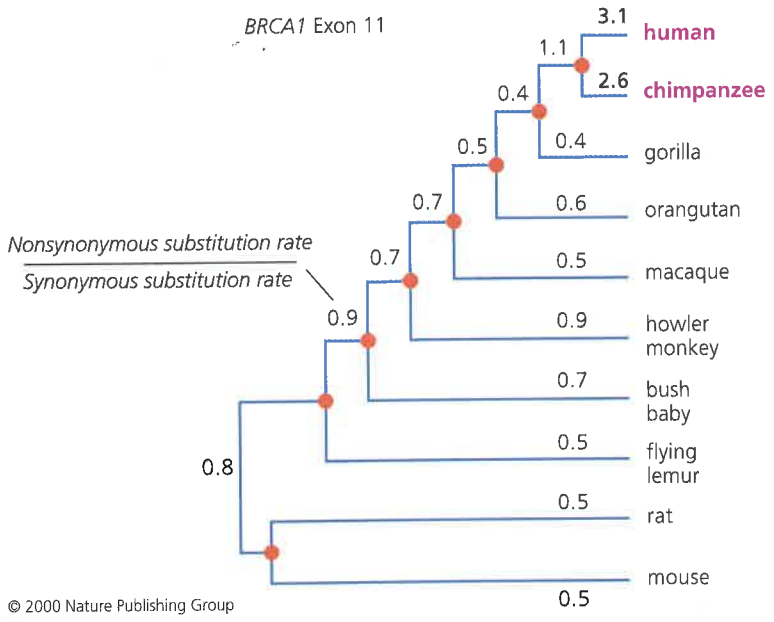

Figure 7.29 Positive selection on the *BRCA1* gene in humans and chimpanzees On most branches of this phylogeny, the ratio of replacement to silent substitution rates is less than one, consistent with neutral evolution. On the branches leading to humans and chimps, however, the ratio is significantly greater than one—consistent with positive selection.

Reprinted by permission from Macmillan Publishers Ltd: Huttley, G. A., E. Easteal, M. C. Southey, et al. 2000. *Nature Genetics* 25: 410–413.

chimps. The selective agent responsible remains unknown, although Pulvers and Huttner (2009) speculate that it may involve brain size.

Comparing Silent and Replacement Changes within and between Species. The research by Hughes and Nei and by Huttley and colleagues provides clear examples of gene segments where neutral substitutions do not predominate. Thanks to the efforts of numerous researchers, many other loci have been found where replacement substitutions outnumber silent substitutions.

Even though the $\frac{d_N}{d_S}$ criterion for detecting positive selection has been useful, Paul Sharp (1997) notes that it is extremely conservative. Replacement substitutions will outnumber silent substitutions only when positive selection has been strong. In a comparison of 363 homologous loci in mice and rats, for example, only one showed an excess of replacement over silent changes. But as Sharp notes (1997, p. 111), "It would be most surprising if this were the only one of these genes that had undergone adaptive changes during the divergence of the two species." Are more sensitive methods for detecting natural selection available?

John McDonald and Martin Kreitman (1991) invented a test for natural selection that is widely used. The McDonald–Kreitman, or MK, test is based on the neutral theory's assertion that all standing variation at both silent sites and replacement sites consists of neutral alleles evolving by drift (see Fay 2011). If this assertion is true, then the ratio of nonsynonymous to synonymous substitutions between closely related species, $\frac{d_N}{d_S}$, should be the same as the ratio of synonymous to nonsynonymous polymorphisms within species, $\frac{p_N}{p_S}$. A **polymorphism** is a locus at which different individuals in a population carry different alleles. Positive selection on nonsynonymous substitutions within species can elevate $\frac{d_N}{d_S}$ above $\frac{p_N}{p_S}$ because beneficial mutations rise quickly to fixation within populations. They thus contribute only briefly to polymorphism, but permanently and cumulatively to interspecific divergence.

Researchers have developed statistical tests for detecting positive selection that are more sensitive than the simple ratio of nonsynonymous to synonymous substitution.

McDonald and Kreitman's initial use of this test compared sequence data from the alcohol dehydrogenase (*Adh*) gene of 12 *Drosophila melanogaster*, 6 *D. simulans*, and 12 *D. yakuba* individuals. *Adh* was an interesting locus to study because fruit flies feed on rotting fruit that may contain toxic concentrations of ethanol, and the alcohol dehydrogenase enzyme catalyzes the conversion of ethanol to a nontoxic product. Because of the enzyme's importance to these species, and because ethanol concentrations vary among food sources, it is reasonable to suspect that the locus is under selection when populations begin exploiting different fruits.

In an attempt to sample as much within-species variation as possible, the individuals chosen for the study were from geographically widespread locations. McDonald and Kreitman aligned the *Adh* sequences from each individual in the study and identified sites where a base differed from the most commonly observed nucleotide, or what is called the consensus sequence. The researchers counted differences as fixed if they were present in all individuals from a particular species, and as polymorphisms if they were present in only some individuals from a particular species. Differences that were fixed in one species and polymorphic in another were counted as polymorphic.

McDonald and Kreitman found that 29% of the differences that were fixed between species were replacement substitutions. Within species, however, only 5% of the polymorphisms in the study represented replacements. Rather than being the same, these ratios show an almost sixfold, and statistically significant, difference ($p = 0.006$). This is strong evidence against the neutral model's prediction.

McDonald and Kreitman's interpretation is that the differences in replacement mutations fixed in different species are selectively advantageous. They suggest that these mutations occurred after *D. melanogaster*, *D. simulans*, and *D. yakuba* had diverged and spread rapidly to fixation due to positive selection in the differing environments occupied by these species.

Using the MK test, natural selection has now been detected in loci from plants, protists, and a variety of animals (Escalante et al. 1998; Purugganan and Suddith 1998). With an extension of the MK test applied to 35 genes in *D. simulans* and *D. yakuba*, Nick Smith and Adam Eyre-Walker (2002) estimated that 45% of all amino acid substitutions between the genomes of the two species were fixed by positive selection. With an extension applied to the genomes of humans and chimpanzees, Carlos Bustamante and colleagues (2005) identified 304 human genes that have evolved under positive selection.

Which Loci Are under Strong Positive Selection? Thanks to studies employing the Hughes and Nei analysis, the MK test, and other strategies, generalizations are beginning to emerge concerning the types of loci where positive natural selection has been particularly strong (Yang and Bielawski 2000; Vallender and Lahn 2004; Nielsen 2005; Nielsen et al. 2005). Replacement substitutions appear to be particularly abundant in loci involved in arms races between pathogens and their hosts (for example, Hughes and Nei 1989), in loci with a role in reproductive conflicts such as sperm competition and egg–sperm interactions (Swanson and Vacquier 1998; Dorus et al. 2004), and in recently duplicated genes that have attained new functions (Zhang et al. 1998). Positive selection has also been detected in genes involved in sex determination, gametogenesis, sensory perception, interactions between symbionts, tumor suppression, and programmed cell death as well as in genes that code for certain enzymes or regulatory proteins.

Positive selection seems to be particularly common in genes involved in biological conflict.

As data accumulate from genome-sequencing projects in closely related species, such as humans and chimpanzees, the number and quality of comparative studies are exploding. Even before the era of genome sequencing began, however, it became clear that silent substitutions, as well as replacement changes, are subject to natural selection.

Selection on "Silent" Mutations

The term *silent mutation* was coined to reflect two aspects of base changes at certain positions of codons: They do not result in a change in the amino acid sequence of the protein product, and they are not exposed to natural selection. The second proposition had to be discarded, however, in the face of data on phenomena known as codon bias, hitchhiking, and background selection. How can mutations that do not alter an amino acid sequence be affected by natural selection?

Direct Selection on Synonymous Mutations: Codon Bias and Other Factors. Most of the 20 amino acids are encoded by more than one codon. We have emphasized that changes among redundant codons do not cause changes in the amino acid sequences of proteins, and we have implied that these silent changes are neutral with respect to fitness. If this were strictly true, we would expect codon usage to be random, and in a given species each codon in a suite of synonymous codons to be present in proportions that reflect the G+C content of the species' genome. But early sequencing studies confirmed that codon usage is highly nonrandom (Figure 7.30). This phenomenon is known as **codon bias.**

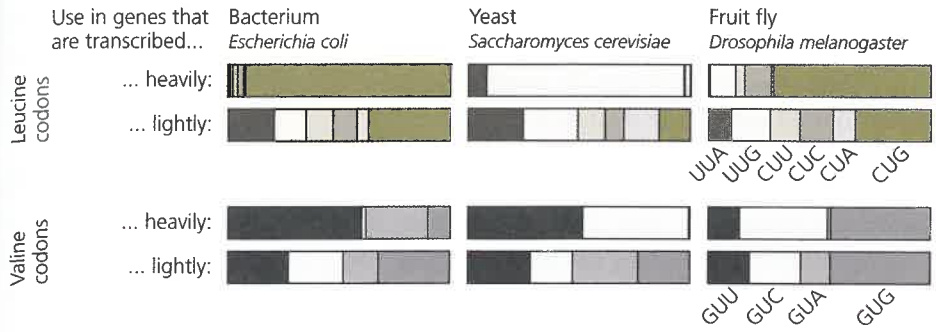

Figure 7.30 Codon bias Bars show relative use of possible codons for two amino acids in genes with different transcription levels. Lightly transcribed genes use all available codons in roughly equal amounts. Heavily transcribed genes tend to use one or two codons to the exclusion of others. Drawn from data in Sharp et al. (1998).

Several important patterns have emerged from studies of codon bias. Codon bias is strongest in highly expressed genes—such as those for the proteins found in ribosomes—and weak to nonexistent in rarely expressed genes. In addition, the suite of codons that are used most frequently correlates strongly with the most abundant species of tRNA in the cell (Figure 7.31).

The leading hypothesis to explain these observations is natural selection for translational efficiency (Sharp and Li 1986; Sharp et al. 1988; Akashi 1994). The logic here is that if a "silent" mutation in a highly expressed gene creates a codon that is rare in the pool of tRNAs, the mutation will be selected against. The selective agent is the speed and accuracy of translation. Speed and accuracy are especially important when the proteins encoded by particular genes are turning over rapidly and the corresponding genes must be transcribed continuously. It is reasonable, then, to observe the strongest codon bias in highly expressed genes.

Codon bias suggests that some synonymous mutations are not selectively neutral.

Selection against certain synonymous mutations represents a form of negative selection; it slows the rate of molecular evolution. As a result, codon bias may

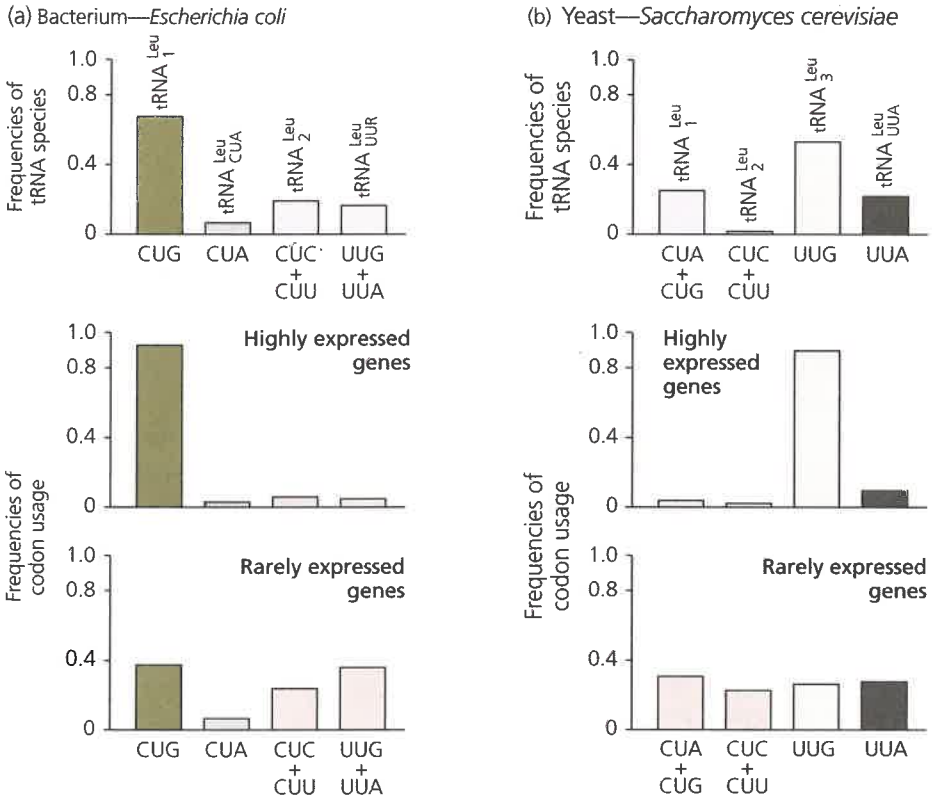

Figure 7.31 Codon bias correlates with the relative frequencies of tRNA species The bar chart in the top row of both (a) and (b) shows the frequencies of four different tRNA species that carry leucine in *E. coli* (a) and the yeast *Saccharomyces cerevisiae* (b). The bar charts in the middle and bottom rows report the frequency of the mRNA codons corresponding to each of these tRNA species in the same organisms. The mRNA codons were measured in two different classes of genes: those that are highly transcribed (middle) and those that are rarely transcribed (bottom). The data show that codon usage correlates strongly with tRNA availability in highly expressed genes, but not at all in rarely expressed genes. Redrawn from Li and Graur (1991).

explain the observation that silent changes do not accumulate as quickly as base changes in pseudogenes. Other synonymous mutations may experience selection as a result of their effects on mRNA stability or exon splicing (see Chamary et al. 2006). The general message here is that not all redundant sequence changes are "silent" with respect to natural selection.

Indirect Effects on Synonymous Mutations: Hitchhiking and Background Selection. Another phenomenon that affects the rate and pattern of change at silent sites is referred to as **hitchhiking,** or a **selective sweep.** Hitchhiking can occur when strong positive selection acts on a particular amino acid change. As a favorable mutation increases in frequency, neutral or even slightly deleterious mutations closely linked to the favored site will increase in frequency along with the beneficial locus. These linked mutations are swept along by selection and may even ride to fixation. Note that this process occurs when only recombination fails to break up the linkage between the hitchhiking sites and the site under selection.

A striking example of hitchhiking happened on the fourth chromosome of fruit flies. The *Drosophila* fourth chromosome is unusual because it shows no recombination. The entire chromosome is inherited like a single gene.

Andrew Berry and colleagues (1991) sequenced a 1.1-kb region of the fourth chromosome in 10 *Drosophila melanogaster* and 9 *D. simulans*. The region includes the introns and exons of a gene that is expressed in fly embryos and called *cubitus interruptus Dominant* (*ciD*). Within it Berry et al. found no differences whatsoever among the *D. melanogaster* individuals surveyed. The entire 1.1 kb of sequence was identical in all 10 individuals. Among the *D. simulans* they found only one base difference. In other words, there was almost no polymorphism in this region. In contrast, when the researchers compared the sequences between the two species, they found 54 substitutions.

Other chromosomes surveyed in the same individuals showed normal amounts of polymorphism. These latter data serve as a control and confirm that the lack of variation in and around the *ciD* locus is not caused by an unusual sample of individuals. Rather, there is something unusual about the fourth chromosome.

Berry et al. suggest that recent selective sweeps cleaned out all or most of the variation on the fourth chromosome in each species. An advantageous mutation anywhere on the fourth chromosome would eliminate all within-species polymorphism as it rose to fixation. New variants, like the one polymorphism observed in the *D. simulans* sampled, will arise only through mutation. In this way, selective sweeps leave a footprint in the genome: a striking lack of polymorphism within linkage groups. Similar footprints have been found in other chromosomal regions where the frequency of recombination is low, including the ZFY locus of the human Y chromosome (Dorit et al. 1995) and a variety of loci in *D. melanogaster* and other flies (for example, see Nurminsky et al. 1998).

Selection at nearby sites can influence the evolutionary fate of synonymous mutations.

Has hitchhiking produced all of these regions of reduced polymorphism? Probably not. Another process, called **background selection,** can produce a similar pattern (Charlesworth et al. 1993). Background selection results from negative selection against deleterious mutations, rather than positive selection for advantageous mutations. Like hitchhiking, it occurs in regions of reduced recombination. The idea here is that selection against deleterious mutations removes closely linked neutral mutations and yields a reduced level of polymorphism.

Although hitchhiking and background selection are not mutually exclusive, their effects can be distinguished in at least some cases. Hitchhiking results in

dramatic reductions in polymorphism as an occasional advantageous mutation quickly sweeps through a population. Background selection causes a slow, steady decrease in polymorphism as frequent deleterious mutations remove individuals from the population.

Status of the Neutral Theory

The neutral theory of molecular evolution explains the clocklike evolution of nucleotide sequences we saw in Figures 7.24 and 7.26. It also explains why silent substitutions outnumber replacement substitutions in most genes, as we saw in Figures 7.24 and 7.25. And the neutral theory serves as a null hypothesis that allows researchers to identify examples of positive selection on nucleotide sequences, as illustrated in Figure 7.29. By all these criteria, the neutral theory of molecular evolution is extraordinarily useful.

As a null hypothesis for detecting positive selection in molecular evolution, the neutral theory has been highly successful.

What about the theory's fundamental claim that the vast majority of nucleotide changes that become fixed in populations are selectively neutral and that molecular evolution is largely due to genetic drift? To assess this claim, we need (1) data for as many substitutions as possible in as many species as possible, and (2) a breakdown of the proportion of substitutions that are neutral versus deleterious versus beneficial. The data we need are accumulating. To assemble the information summarized in Figure 7.32, Justin Fay combed the literature to compile estimates of α, the fraction of amino acid substitutions driven by positive selection. He included data on 38 species for which multiple genes have been studied.

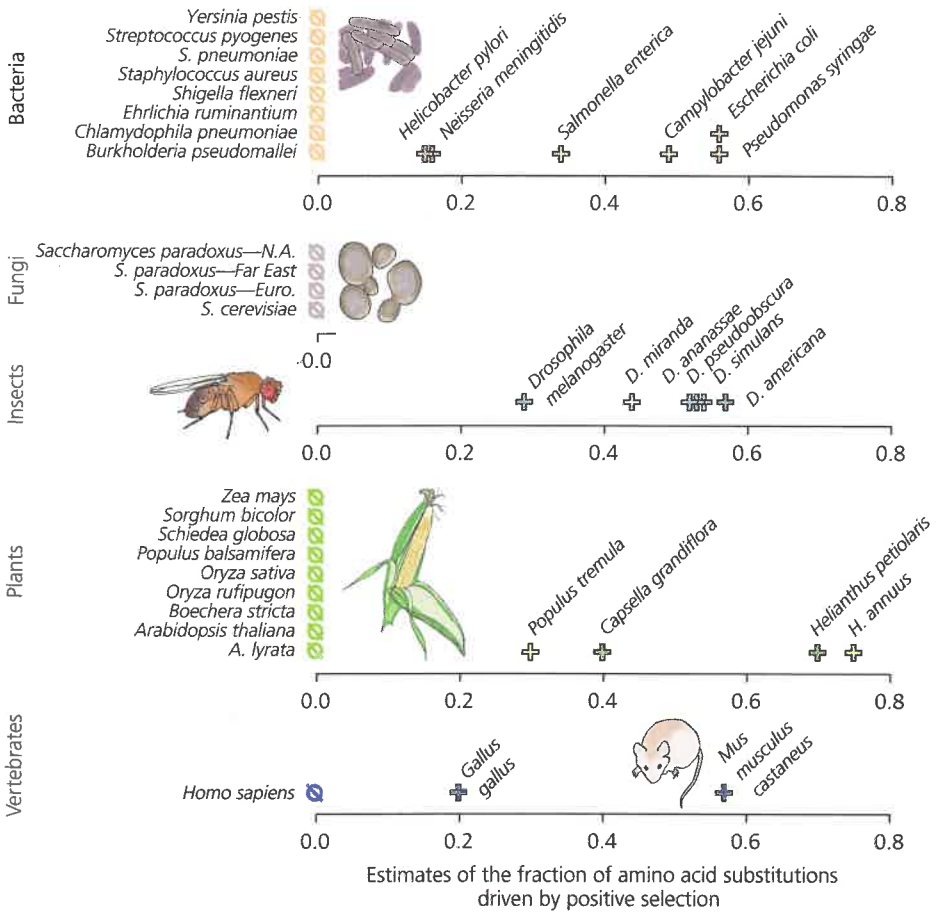

Figure 7.32 Estimates for different species of the fraction of amino acid substitutions driven by positive selection Null signs indicate a lack of statistically significant evidence, based on the McDonald–Kreitman test, for positive selection. Light plus signs indicate that there is conflicting evidence from different studies. The agents of selection are generally unknown. Data from Fay (2011).

Alpha can be estimated from a McDonald–Kreitman test as the elevation of $\frac{d_N}{d_S}$ above $\frac{p_N}{p_S}$:

$$\alpha = 1 - \frac{\left(\frac{p_N}{p_S}\right)}{\left(\frac{d_N}{d_S}\right)}$$

Plus signs appear in the graph only for those species for which the McDonald–Kreitman test gave statistically significant evidence of positive selection.

At first glance, the data appear to refute—at least for some species—the neutral theory's claim that selectively neutral mutations that rise to fixation by drift vastly outnumber beneficial mutations that rise to fixation by natural selection. Fay argues, however, that it is too early to draw such a conclusion. The McDonald–Kreitman test does not distinguish between positive selection and other mechanisms that can lead to elevated levels of nonsynonymous divergence between species (see also Hughes 2007; Nei et al. 2010). One alternative is reduced population size, which can lead to fixation by drift of mildly deleterious mutations. Another is hitchhiking. If an unknown number of linked deleterious substitutions ride to fixation with a single positively selected one, the true proportion of substitutions driven by positive selection is obscured.

Data are now accumulating that will allow researchers to evaluate the neutral theory's claim that most molecular evolution is dominated by negative selection and drift. For now, the issue is undecided.

The jury is still out on the neutral theory's fundamental claim.

Coalescence

Before closing our discussion of genetic drift and molecular evolution, we want to mention another area of research in which sequence data and the null model of genetic drift are being put to productive use. This is the study of coalescence. Here we consider coalescence as a tool for estimating effective population size, although it has a great variety of other applications.

Coalescence Defined

Figure 7.21 showed an evolving population in cartoon form. New alleles arose by mutation and became more common over time as each copy propagated additional copies into future generations. Imagine what we would see if we could reverse the flow of time and watch the population de-evolve. The blue allele would become rarer as descendent copies merged into their common ancestors. So, too, would the dark green allele. The blue allele would disappear as the original copy merged into the dark green copy it sprang from. Then the dark green allele would disappear as the original copy merged into its light green progenitor.

If we could run the movie of molecular evolution backward, we would see alleles become less divergent and eventually merge into their common ancestral sequence. This process is called coalescence.

Now imagine that we have a sample of real alleles from a population of organisms. Each represents an unbroken lineage of copies descended from copies in ever earlier generations. If we could trace these lineages back in time, we would see them merge until only one lineage, the last common ancestor of our sampled alleles, remained. The merging of genealogical lineages as we trace allele copies backward in time is called **coalescence.**

The term was coined by John Kingman (see Kingman 2000), who found a way to simulate the coalescence of alleles in a population evolving backward in time by genetic drift. Among his method's virtues is that it requires no information about the rest of the population other than its size (see Felsenstein 2004). The result is an evolutionary tree of genes—a **gene tree** or **gene genealogy.**

Mathematical descriptions of coalescence provide an efficient means of simulating evolution by genetic drift.

Figure 7.33a shows several gene genealogies resulting from simulated coalescence of seven alleles in populations of 1,000 and 5,000 individuals. Notice first that every one of the simulated gene trees is unique. We are modeling genetic

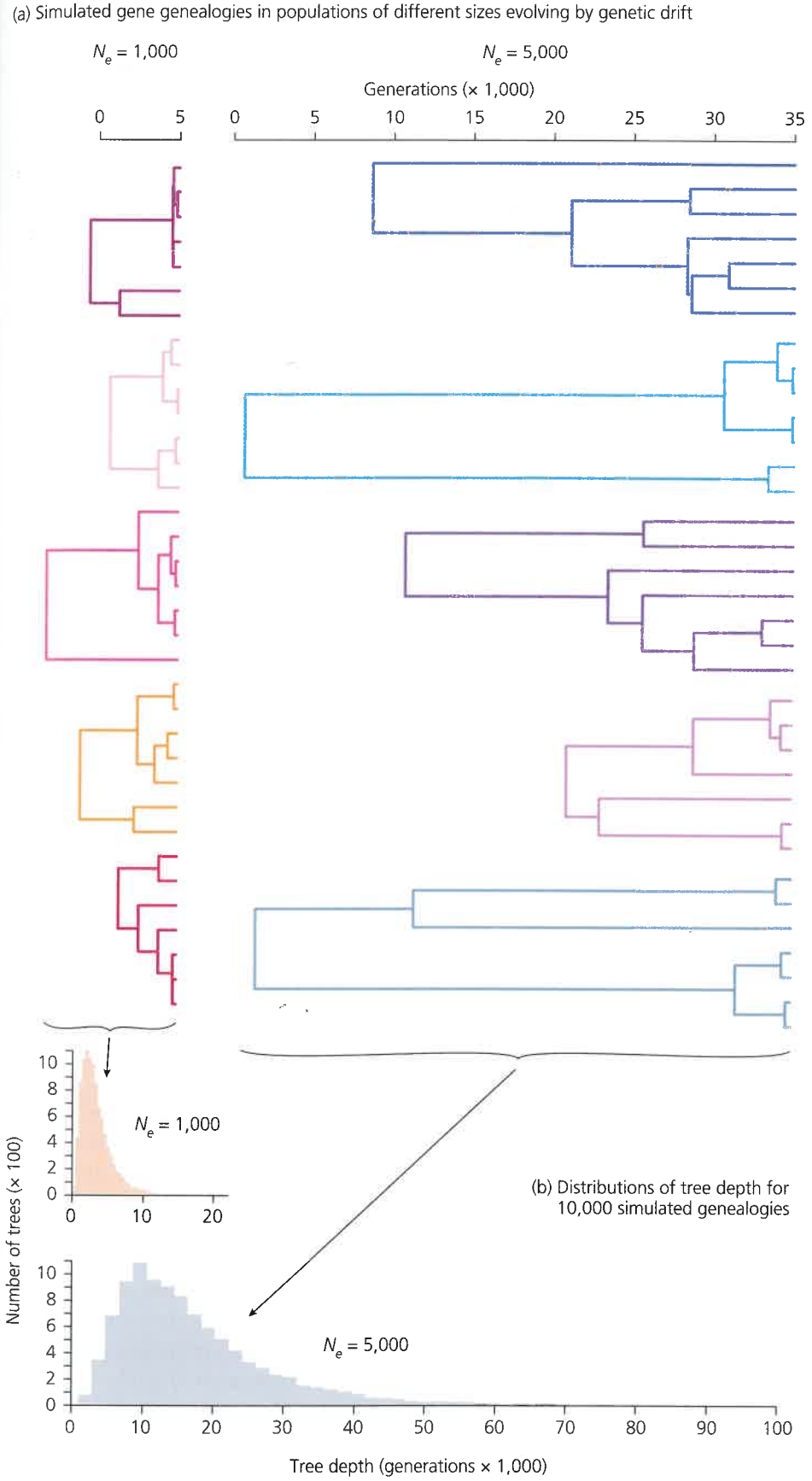

Figure 7.33 Gene genealogies produced by simulation of coalescence (a) Examples of genealogies produced by simulating genetic drift running backward in time. The five trees in the column on the left are examples of the results from simulating the coalescence of seven alleles in a population of 1,000 individuals. The five trees in the column on the right are examples of the results from simulating the coalescence of seven alleles from a population of 5,000 individuals. All trees are drawn to the same scale. (b) Distributions of tree depth, or time (in generations) back to the most recent common ancestor, for 10,000 simulated trees from a population of 1,000 and a population of 5,000. Simulations performed and distributions generated by *Mesquite* (Maddison 2011; Maddison and Maddison 2011). Trees drawn by PHYLIP (Felsenstein 2009a).

drift, so the differences among trees are due to chance events. Second, notice that the coalescent trees for alleles in populations of 5,000 (right column) tend to be deeper than the trees for alleles in populations of 1,000 (left column). We have to travel further back in time to find the common ancestor of alleles sampled from a large population. This makes sense. Randomly chosen individuals from a large population are likely to be more distantly related than randomly chosen individuals from a small population (Kuhner 2009).

Figure 7.33b documents this observation in more detail. It shows the distribution of tree depths among 10,000 simulated coalescent trees for each population size. Like the trees, the distributions are drawn on the same scale. The distributions overlap, but they nonetheless suggest a method we could use to estimate the effective size of a real population.

Coalescence Applied

Imagine we had a sample of allelic sequences from seven randomly chosen individuals from a population of unknown size. Imagine further that we knew the true genealogy of the seven alleles and its depth (which would require that we also knew the mutation rate). Finally, imagine that the depth was, say, 10,000 generations. Comparing the two distributions in Figure 7.33b shows that we could not infer the size of our population with certainty. But we could conclude that 5,000 is a much better guess than 1,000.

Of course, we do not know the true genealogy of our seven alleles, nor do we know the mutation rate. We could simply estimate the tree (with methods discussed in Chapter 4) and the mutation rate (using data discussed in Chapter 5). It would be tempting to make these estimations and treat the results as true for purposes of comparison with our simulated gene trees. The problem with doing so is that it ignores the uncertainty associated with estimation (Felsenstein 2009b).

A better approach is to use techniques related to the likelihood and Bayesian methods for inferring phylogenies that we discussed earlier (Chapter 4). We start with a model of molecular evolution, which in the present case would include parameters for the mutation rate and population size. We then use a type of software called a coalescent genealogy sampler to search the universe of possible gene trees and parameter values (see Kuhner 2008). For each combination of tree and parameter values it considers, the software calculates a metric reflecting how good an explanation that particular model offers for our data. At the end of the search, which is long and computationally demanding, the software can give us a range of plausible values for the size of the population our set of alleles came from. We can increase the accuracy of our estimate by including sequence data for sets of alleles at as many independent loci as possible (Felsenstein 2006).

Coalescence models can be fit to data, yielding estimates for parameters such as population size.

Elizabeth Alter and colleagues (2007) used the approach we have outlined to estimate the effective population size of the gray whale (*Eschrichtius robustus*). The researchers analyzed data for several dozen alleles at each of 10 independent loci. Coalescence analysis indicated that the whales' genetic diversity was consistent with deeper gene genealogies, and thus a much larger effective population size than would be expected from their current census population size. The best explanation is the obvious one. Commercial whaling in the recent past drastically reduced the population size and, despite claims to the contrary, the whales' numbers have yet to return to their pre-whaling abundance.

For more on the coalescent theory and its applications, see Felsenstein (2004). For a book-length treatment, see Wakeley (2009).