CHAPTER THREE

Properties of DNA Sequence Loci: Part II

In Chapter 2, we reviewed the composition of eukaryotic genomes and aspects of molecular evolution in order to lay the groundwork for our understanding of the important properties of DNA sequence loci. Some locus properties may be beneficial to one type of phylogenomic study or problematic to another. For example, to the worker who is interested in studying the evolutionary diversification of a gene family such as the mitochondrial pseudogenes of a particular genome, the property of being a duplicated or "multicopy" gene will be of intrinsic interest. In contrast, to the worker who is interested in using mitochondrial genes as evolutionary markers for purposes of tracking organismal diversification or for identifying species (i.e., DNA barcoding), the existence of mitochondrial pseudogenes is a major source of concern and therefore special precautionary steps must be taken to avoid potential problems.

We will now continue our discussion of loci properties by focusing on several commonly made assumptions in phylogenomic analyses particularly those pertaining to tree of life studies. These assumptions include that each locus (1) exists as a single-copy in the genome; (2) has not been under the direct or indirect influence of natural selection (i.e., is selectively neutral); (3) has a gene tree that is independent of the gene trees of other sampled loci; (4) has not experienced a recombination event since the most recent common ancestor of the sampled sequences (i.e., all sites have the same gene tree); (5) has a constant substitution rate among lineages (i.e., molecular clock); and (6) has been chosen randomly from the genome (i.e., no ascertainment bias). Following this discussion, we will briefly revisit the conceptual meanings of important terms such as locus, gene, allele, gene copy, and haplotype before we review the actual types of loci used in phylogenomic studies.

3.1 SIX ASSUMPTIONS ABOUT DNA SEQUENCE LOCI IN PHYLOGENOMIC STUDIES

3.1.1 Assumption 1: Loci Are Single-Copy in the Genome

In Chapter 2, we saw that certain types of loci have been duplicated one or more times and thus exist as multiple copies in genomes. These include tandemly repeated loci, transposable elements, *numts*, and perhaps others. While the existence of these gene families is of great interest to researchers interested in molecular evolution, these multicopy genes can be a nightmare for the organismal biologist. This is because gene trees inferred from different copies of a homologous gene reveal the history of gene duplication events, but they obfuscate the phylogenetic history of the study organisms even if the gene tree is correct (Moritz and Hillis 1996). Thus, the single-copy locus assumption is fundamental to phylogenomics studies of the tree of life (Moritz and Hillis 1996; Delsuc et al. 2005; Philippe et al. 2005; Philippe and Blanchette 2007).

The term *homology* refers to common ancestry and thus DNA sequences displaying similar sequences are presumed to be homologous (Moritz and Hillis 1996). However, because homology must be inferred it is important to realize that similarity does not equate with homology (Moritz and Hillis 1996 and references therein). Nonetheless, in practice researchers typically use software to align sequences based on the site-by-site

similarities or what is referred to as *positional homology* (Moritz and Hillis 1996). Obtaining a satisfactory alignment is usually trivial when aligning either recently diverged sequences or protein-coding sequences, whereas it can be much more challenging or impossible if numerous insertion and deletion mutations have occurred such as in the loop structures of ribosomal RNA genes or in introns between anciently diverged organisms (Moritz and Hillis 1996).

For molecular studies, it is necessary to distinguish several different homology concepts owing to the nature of how molecular sequences can evolve particularly with respect to the processes of gene duplication and horizontal gene transfer (HGT). Fitch (1970) first brought this topic to the attention of molecular phylogeneticists when he pointed out that two proteins (or their underlying DNA sequences) can be similar either because (A) they descended with divergence from a gene in a common ancestor via speciation (viz. Darwin's descent with modification) or (B) because the two sequences descended with convergence from two separate ancestors via gene duplication (Fitch 1970). To distinguish these forms of "homology," Fitch (1970) coined the term *orthology* to characterize similar sequences in scenario (A) and *paralogy* to describe similar sequences following scenario (B). Thus, if a researcher uses a particular genomic locus that has not been duplicated, then all sequences obtained from other individuals (or species) will likely represent orthologous copies. On the other hand, if a researcher interested in organismal phylogeny obtains a mixed sample of "homologous" sequences that includes orthologous and paralogous sequences, then the inferred gene tree for these sequences will almost certainly provide nonsensical or misleading results.

As an example of how paralogous DNA sequences can cause problems in organismal phylogenomic studies let's look at the well-studied problem of mitochondrial pseudogenes. *Numts* are especially worrisome when PCR is used to obtain the target mitochondrial sequences because there are at least two adverse scenarios possible. In the first scenario, the target mitochondrial gene and *numt* are coamplified in PCR (Zhang and Hewitt 1996; Sorenson and Quinn 1998; Bensasson et al. 2001). Unless the two amplification products have the identical sequence, which could occur if the *numt* represents a recent primary integration, then it is likely that the resulting Sanger sequence will be of poor quality if not ruined (we will see why in Chapter 6). In the second scenario, a *numt* is preferentially amplified over the target mitochondrial gene (Smith et al. 1992; Collura and Stewart 1995), which can lead to datasets consisting of all sequences from one *numt* locus, mixtures of *numt* paralogues, or mixtures of *numts* and the target mitochondrial locus.

Numt-contaminated datasets can cause a variety of problems ranging from obtaining poor quality DNA sequences to spurious interpretations of phylogenomic results. Specifically, results based on *numt*-contaminated datasets can mislead researchers interested in using the gene tree as a proxy of the species tree (Zhang and Hewitt 1996; Bensasson et al. 2001). *Numts* can also be problematic in studies of mitochondrial diseases in which discovery of "novel" mutations in *numts* might be falsely linked to diseases (Hazkani-Covo et al. 2010). Also because substitution rates for *numt* sites are 10-fold slower than their mitochondrial counterparts (Brown et al. 1979), the accidental analysis of *numt* sequences could lead to grossly underestimated evolutionary distances. This is especially a concern for DNA barcoding studies because genetic distances are often used to identify cryptic species and thus spurious distance estimates may lead to underestimates of biodiversity. *Numts* can also confound phylogeographic analyses because their effective population sizes are four times larger than mitochondrial genes (Zhang and Hewitt 1996; Zink and Barrowclough 2008). Sorenson and Quinn (1998) reviewed the *numt* phenomenon in birds and suggested a number of laboratory strategies for preventing the unintended sequencing of *numts* as well as methods for identifying putative *numts* in existing DNA sequence datasets.

Not all types of paralogous loci will likely cause problems in tree of life studies. Some tandemly repeated loci such as nuclear rDNA arrays, because of their proximity to each other along the chromosome, usually evolve via *concerted evolution* (Zimmer et al. 1980; Hillis and Dixon 1991; Moritz and Hillis 1996). Thus, as the tandemly arrayed loci evolve together, the conceptual distinction between the original orthologous versus adjacent paralogous copies is largely erased (Moritz and Hillis 1996). For example, Hillis and Davis (1986) analyzed variation within tandem arrays of nuclear ribosomal DNA genes in a 50 million year old group of ranid frogs in order to

examine the evolution of the genes and to infer a molecular phylogeny.

Lastly, there is yet another form of homology that we need to define for cases in which genes are transferred horizontally from one species' genome to another. This is frequently referred to as HGT or lateral gene transfer. The term *xenology* was coined to describe genes that are transferred from the genome of one species to the genome of another species (Gray and Fitch 1983). HGT has played a significant role in the evolution of prokaryotic and eukaryotic genomes. For further discussion and perspectives about the various homology concepts, the reader should consult the article by Mindell and Meyer (2001).

3.1.2 Assumption 2: Loci Are Selectively Neutral

Neutral coalescent theory-based approaches to estimating phylogeographic and species tree parameters typically require that each locus has not been directly or indirectly influenced by natural selection (e.g., Hudson and Coyne 2002). In other words, neutral sites evolve only via mutation and genetic drift (e.g., Fu and Li 1993; Wakeley 2009). This assumption is often stated as loci are assumed to be *selectively neutral*. By harvesting large numbers of selectively neutral loci from genomes, researchers can take advantage of coalescent-based methods for inferring the evolutionary history of populations and species. However, this discussion about selectively neutral DNA presumes that genomes contain functionless DNA, an idea that has been controversial for a long time.

3.1.2.1 Does "Junk DNA" Exist?

During the 1960s, two different but related lines of thought were aired among geneticists, which set into motion a long-standing debate of importance to the field of molecular evolution. First, researchers postulated the existence of functionless or "junk" DNA (Graur et al. 2015 and references therein). Ohno (1972) later defined junk DNA as DNA on which natural selection does not operate. In other words, junk DNA does not presently have any advantageous or deleterious effect on the fitness of its carrier. Secondly, the introduction of the neutral theory of molecular evolution by Kimura (1968) and King and Jukes (1969) represented the other important development (see Gillespie 2004 for review; Eisen 2012). The empirical results of the first genetic electrophoresis study by Lewontin and Hubby (1966) revealed much higher levels of genetic (protein) variation in natural *Drosophila pseudoobscura* populations than would have been expected due to prevailing thoughts at that time about molecular evolution—namely, that most mutations were harmful and therefore would be promptly eliminated from populations via purifying natural selection (a.k.a. the "Darwinian" evolution viewpoint; Yang 2006; Eisen 2012). But the existence of such high levels of genetic variability in *Drosophila* populations demanded an explanation and the thoughts about "neutral evolution" in the papers by Kimura and by King and Jukes helped accommodate these surprising results (Yang 2006; Eisen 2012). This opposing viewpoint was encapsulated in the title of the King and Jukes (1969) paper: *Non-Darwinian Evolution*.

What exactly is junk DNA and does it really exist? Recall from Chapter 2 that intergenic DNA is largely comprised of DNA with no known function. Comparative genome studies provide evidence supporting the hypothesis that most intergenic DNA is without biological function. Meader et al. (2010) and Ponting and Hardison (2011) suggested that about 85% of the human genome is comprised of nonfunctional DNA. The results of another study suggest that an even higher fraction—95% of the human genome—is comprised of nonfunctional DNA (Lindblad-Toh et al. 2011). However, this idea that any genome could contain a large amount of functionless DNA has remained controversial and the human genome, in particular, has been the primary battleground (ENCODE Project Consortium 2012; Hurtley 2012; Pennisi 2012; Doolittle 2013; Graur et al. 2013; Palazzo and Gregory 2014; Graur et al. 2015; Palazzo and Lee 2015).

For many researchers there is no longer any question as to whether junk DNA exists. Recent papers (Doolittle 2013; Graur et al. 2013; Palazzo and Gregory 2014; Palazzo and Lee 2015) have provided strong arguments in favor of junk DNA's existence. The junk DNA controversy has arisen because researchers conflated different meanings of the word "function." For example, ENCODE researchers studying the human genome used data indicating the existence of a biochemical process (e.g., transcription) as de facto evidence for those genomic sequences being "functional" without consideration if

those sequences may have been maintained by natural selection (Eisen 2012; Graur et al. 2013). This does not mean that *some* of the currently classified nonfunctional DNA in genomes won't eventually be shown to have an essential function maintained by purifying selection (Eisen 2012; Graur et al. 2013). However, at least in vertebrate genomes, there are vast tracts of intergenic DNA that are devoid of any evolutionary conservatism (Ovcharenko et al. 2005).

Nobel Prize winner Sydney Brenner offered perspective by way of a metaphor as to why we should be more comfortable with the concepts of functional and nonfunctional DNA. According to Brenner (1998), in the broadest sense, nonfunctional DNA is simply *rubbish DNA*. Within this category, Brenner described two subcategories defined on the basis of their evolutionary effects. The first category is *junk DNA*, which is DNA that does not help or harm the organism but could be useful in the future (viz. Ohno 1972). Brenner used the metaphor of rubbish in someone's garage—if the rubbish doesn't cause any problems for the homeowner, then such junk can persist in that space indefinitely. The second category is *garbage DNA*, which constitutes DNA that adversely affects the individual's fitness. In this case, natural selection would sooner or later eliminate such DNA from the genome much in the way that someone discards hazardous garbage stored in their garage (e.g., toxic chemicals).

Graur et al. (2015) proposed an evolutionary classification scheme for DNA elements that is based in part on Brenner's outline. In this scheme, genomic elements are first classified as being "functional" or "rubbish" depending on whether natural selection has maintained those elements or not. Thus, functional DNA includes DNA that has been selected for a particular function, whereas rubbish DNA has not been selected for a function. Functional DNA, in turn, is comprised of two subcategories called "literal DNA" and "indifferent DNA." Literal DNA is defined as DNA whose actual nucleotide sequence has largely been maintained by selection, while indifferent DNA is defined as DNA that is also under the effects of selection but only for its presence in the genome. Rubbish DNA is also comprised of two subcategories: "junk DNA" is neutral with respect to the fitness of the organism, whereas "garbage DNA" diminishes the fitness of the individual.

Thus, until demonstrated otherwise, it is safe to assume that functionless DNA does exist in genomes. As we will see in Chapter 8, this junk DNA component of genomes represents a vast and largely untapped source for DNA sequence loci that are especially well suited to some types of phylogenomic analyses (e.g., coalescent-based studies of species trees).

3.1.2.2 *The Neutrality Assumption and the Indirect Effects of Natural Selection*

Some genomic sites such as those within introns can tolerate base substitutions and indels, which probably do not affect the fitness of the individual. Therefore, in a sense, they are selectively neutral mutations. Although sites not maintained by selection may appear to be "neutral" sites, this does not necessarily mean that all of these sites will meet the neutrality assumption in phylogenomic studies. The assumption of selective neutrality requires each locus to be free not only of direct natural selection but also from the indirect effects as well. In other words, a site (or locus) that meets this assumption will have a genealogy (gene tree) that has not been distorted by any form of natural selection. However, purifying or positive selection acting on functionally important sites can influence the genealogical histories of linked nonfunctional sites; when this occurs, these gene trees will not reflect the gene tree shapes predicted by neutral theory (Kaplan et al. 1989; Charlesworth et al. 1993, 1995; Charlesworth 2012).

Earlier in this chapter we saw that a large fraction of some genomes (e.g., human) is comprised of nonfunctional DNA. Does any of this nonfunctional DNA consist of sites that are also free of the *indirect* effects of selection and thus represent selectively neutral DNA? Lee and Edwards (2008) compared the nucleotide diversities (π), which is a measure of genetic diversity, of 29 anonymous loci versus six intron loci and one mitochondrial (ND2) gene. Anonymous loci are from random genomic locations and thus many of them are expected to be located far from functional genomic elements, whereas introns contain some functional sites and are located adjacent to exons. In Figure 3.1, we see that the genetic diversities of anonymous loci tend to be higher than for the introns or ND2 coding region. These results are compelling because it suggests that genomes may contain many loci that are effectively free of all

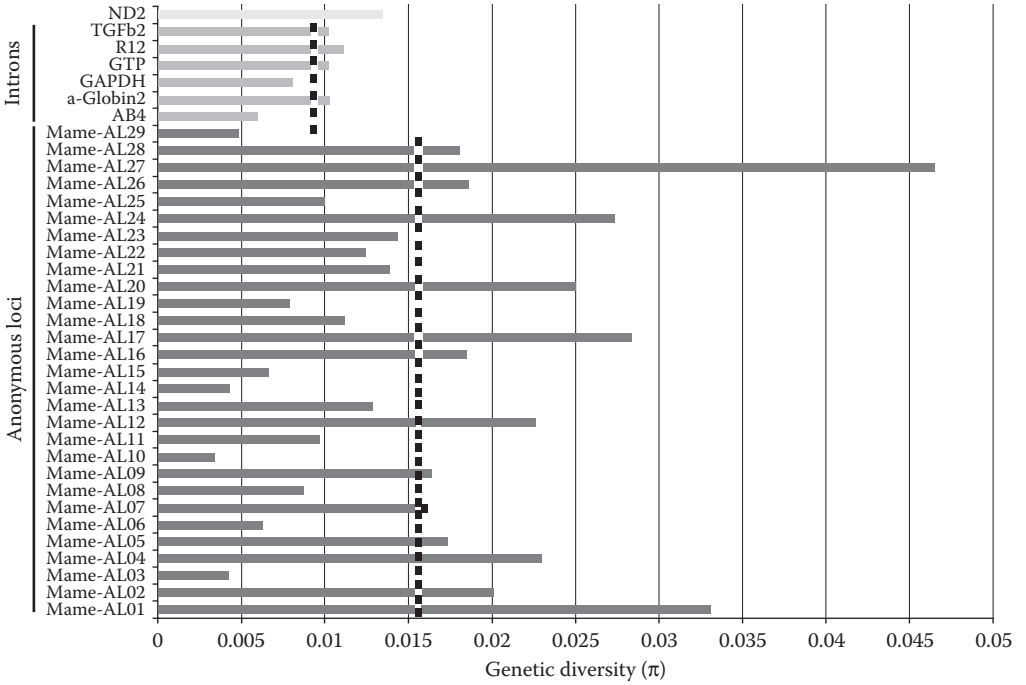

Figure 3.1. Comparison of genetic diversities among different types of DNA sequence loci (Lee and Edwards 2008; fig. 2). The legend reads, "Genetic diversity (π) across 29 nuclear markers, six introns, and ND2 region. Dotted vertical bars represent the mean genetic diversity of introns and nuclear anonymous markers." (Reprinted from Lee, J. Y. and S. V. Edwards. 2008. *Evolution* 62:3117–3134. With permission.)

forms of natural selection and hence best meet the assumption of neutrality. Why are these anonymous loci exhibiting higher levels of genetic diversity than the introns and ND2 gene?

A substantial body of theory has built up in the last few decades, mainly based on studies of the *Drosophila* genome, which shows how genomic sites maintained by selection can reduce or increase the genetic diversity (relative to neutral theory expectations) of linked sites not directly under selection (Felsenstein 1974; Maynard-Smith and Haigh 1974; Thomson 1977; Kaplan et al. 1989; Begun and Aquadro 1992; Charlesworth et al. 1993, 1995; Wiehe and Stephan 1993; Hudson and Kaplan 1995; Innan and Stephan 2003; Stephan 2010; Charlesworth 2012). Although work in this area was underway in the 1970s, later empirical findings by researchers greatly stimulated further elaboration of theoretical and empirical work in the 1990s (Stephan 2010; Charlesworth 2012). One of the key empirical discoveries was that genome-wide variation in theta (a measure of genetic diversity) within the *Drosophila melanogaster* genome was correlated with local recombination rates (Begun and Aquadro 1992; Hudson and Kaplan 1995). This genetic pattern, which is illustrated in Figure 3.2, shows that the genetic diversities of 15 loci in the *D. melanogaster* genome are strongly associated with local recombination rates (see also Figure 1 in Begun and Aquadro 1992 and the discussion by Sella et al. 2009). As most functional genomic loci are presumably under purifying selection, we would predict that selection targeting functional sites should reduce the genetic diversity of other nearby sites simply by virtue of their close linkage. In view of this, the empirical results in Figure 3.2 suggests that recombination has the ability to dampen the effects of indirect selection on other sites and that this effect increases with higher recombination rates. As we will soon see, physical distances and effective population sizes also play important roles in diminishing the effects of indirect selection on linked nonfunctional loci. Similar empirical findings have been documented in other organisms (see Innan and Stephan 2003 for review). We will review the leading explanations for this pattern and discuss the implications for phylogenomic loci.

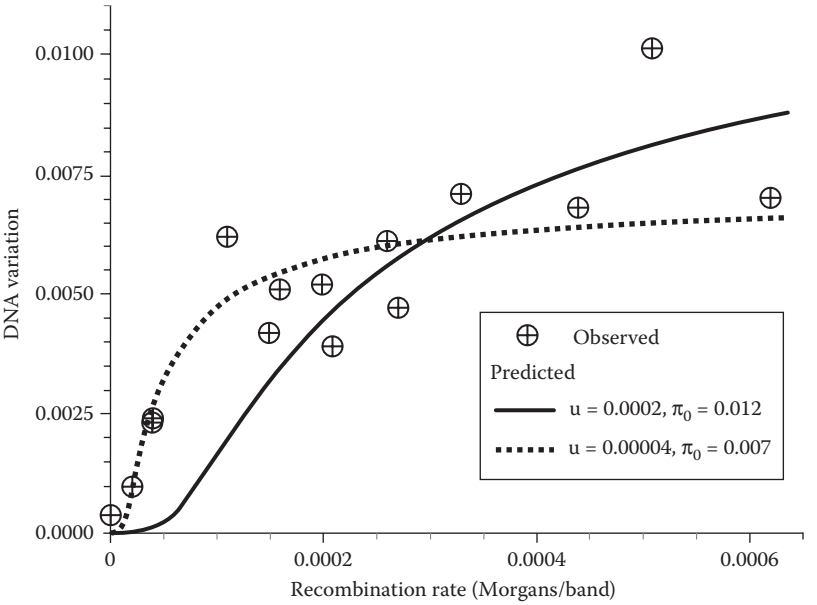

Figure 3.2. DNA variation as a function of local recombination rates in *D. melanogaster* (*Hudson and Kaplan* 1995; *fig.* 2). The legend reads, "Observed and predicted levels of DNA variation as a function of local recombination rates. The observed levels of DNA variation are estimates of theta for 15 loci on the third chromosome of *D. melanogaster* obtained from E. Kindahl and C. Aquadro (personal communication). These loci are, from left to right on the figure, *Lsply*, *Pc*, *Antp*, *Gld*, *Ubx*, *tra*, *fz*, *Mlc2*, *ry*, *Sod*, *Tl*, *Rh3*, *Est6*, *E(spl)*, and *Hsp26*. The local recombination rates for these loci were also provided by E. Kindahl and C. Aquadro. The predicted levels of variation are obtained with (9) with the parameter values indicated on the figure." (Reprinted from Hudson, R. R. and N. L. Kaplan. 1995. *Genetics* 141:1605–1617. With permission.)

The phenomenon of *genetic hitch-hiking* is a form of indirect or correlated selection in which a beneficial mutation quickly spreads throughout the population, which results in a reduction of the genetic diversity at that locus as well as at closely linked sites (Maynard-Smith and Haigh 1974; Thomson 1977; Kaplan et al. 1989; Stephan et al. 1992; Wiehe and Stephan 1993). This phenomenon is also called *selective sweeps* (e.g., Wakeley 2009). Another form of indirect selection called *background selection* occurs when purifying selection acts directly against deleterious mutations, which has the effect of diminishing the genetic diversity of that locus and closely linked sites (Charlesworth et al. 1993, 1995; Hudson and Kaplan 1995; Charlesworth 2012). Felsenstein (1974) had previously described this phenomenon as the *Hill-Robertson Effect* (Charlesworth 2012). *Balancing selection* can also influence the genetic diversity of linked sites. However, in contrast to hitch-hiking, this mode of selection maintains a greater amount of allelic diversity at closely linked sites than predicted by neutral theory (Kaplan et al. 1989). The MHC genes in vertebrates are a classic example of genes influenced by balancing selection and they can exhibit extraordinary levels of genetic diversity (e.g., 200 alleles; Hess and Edwards 2002; Balakrishnan et al. 2010).

The amount of influence that a selected site will have on nonfunctional sites located on the same chromosome is determined by the strength of the linkage relationship. This linkage can be thought of as a correlation in which sites that are tightly linked together have genealogies that are highly correlated with each other. For example, sites within a short locus are often linked to the extent that they all have identical gene trees. This is actually a common phylogenomic assumption, which we will discuss in detail later in this chapter. In contrast, two sites that are weakly linked together such as those separated by large physical distances along the chromosome (on the order of hundreds or thousands of kb) could have uncorrelated genealogies depending on other factors (see below). In the context of indirect selection, if a site that is under selection is tightly linked to a nonfunctional site, the effect will then be stronger because the two sites will remain associated for longer amounts of

time (Santiago and Caballero 1998). On the other hand, if the rates of local recombination and outcrossing are high, then the effect declines rapidly with physical distance (Charlesworth et al. 1993; Hudson and Kaplan 1995).

Physical distance alone is an inadequate measure for characterizing the linkage relationship between two loci when we are concerned about the property of genealogical independence. First, it does not consider the fact that local recombination rates may vary across the genome (e.g., Reich et al. 2001; McVean et al. 2004; Hartl and Jones 2006). Secondly, a physical distance measure does not account for the effective population size, which may also vary across the genome (e.g., Hudson and Kaplan 1995). Physical distance, local recombination rates, and effective population sizes all play a role in determining the strength of association between two loci and their gene trees (Kaplan et al. 1989; Charlesworth et al. 1993).

The *recombinational distance* or Cm ($= 2N_e cm$) combines these factors—effective population size (N_e), per generation recombination rate (c), and physical distance in bp between a selected site and a distant site not directly under selection (m)—into a single measure that can gauge the degree of genealogical independence between two genomic sites (Kaplan et al. 1989). How large does the recombinational distance need to be in order for a nonfunctional site to escape the effects of selection? Kaplan et al. (1989) defined M as the minimum recombinational distance in which a selected site can influence a distant site that is not directly under selection. In other words, when the physical distance between sites is $\geq M/C$ bp, then the two sites will be independent of each other (Kaplan et al. 1989).

Unfortunately, the recombinational distance is more of conceptual than practical value for us at this time. For example, from this quantity we can see that the indirect effects of selection on a distant nonfunctional site (or locus) can be weakened with large population sizes, high rates of crossing over, or large physical distances (Kaplan et al. 1989; Charlesworth et al. 1993). The recombinational "distance" reminds us that population size also plays a large role (Hudson and Kaplan 1995). If a linkage map is available for a species, then a rough distance measure consisting of map units might be used *in lieu* of M (or M/C). For example, Hudson and Kaplan (1995) suggested that sites separated by "a few map units" in the *D. melanogaster* genome would likely not influence each other, at least via background selection. Costa et al. (2016) discuss a genealogically-based approach for specifying threshold distances between known functional elements and distantly located nonfunctional elements in order to identify neutral loci candidates.

3.1.3 Assumption 3: Sampled Loci Have Independent Gene Trees

A common assumption of phylogeographic and species tree analyses that use multilocus coalescent methods holds that all sampled loci (or individual sites) have *genealogically independent histories* or, more simply, *independent gene trees* (Jennings 2016). Such genealogically independent loci are often referred to as "independent loci" or "unlinked loci." The property of independence among sampled loci is important from a statistical point of view because it means that individual loci and hence gene trees inferred from them—can be considered *true replicates* that depict the ancestry of a genome (Edwards and Beerli 2000; Arbogast et al. 2002; Wakeley 2009). Thus, by increasing the numbers of independent loci in a sample, a researcher can obtain more accurate and precise parameter estimates than from datasets with fewer loci (Pluzhnikov and Donnelly 1996; Jennings and Edwards 2005; Felsenstein 2006; Lee and Edwards 2008; Smith et al. 2013; Costa et al. 2016). We will now consider a hypothetical example for a common type of phylogeographic analysis that shows the benefits of many independent loci.

Let's say you want to test the hypothesis that two closely related species diverged from each other during the Pleistocene (i.e., between 0.01 and 2 million years ago). You plan to use two different datasets, which are analyzed separately: one dataset has 12 mitochondrial loci while the other has 12 independent nuclear genes. Let's further assume the resulting two divergence estimates are the following: mtDNA divergence is 3 million years ago, whereas the nuclear estimate is 1.5 million years ago. Given the discrepancy between the two divergence time estimates, which one would you prefer? You should prefer the nuclear dataset results for two reasons. First, a sample of independent nuclear loci permits you to use multilocus coalescent-based methods, which more accurately estimate population divergence times than single locus estimates (Edwards and Beerli 2000). Recall from Chapter 2 that the mitochondrial genome effectively does not recombine therefore all of its

sites can be represented by a single gene tree (e.g., Wilson et al. 1985). Secondly, the nuclear gene estimate will be bracketed by confidence intervals, whereas such confidence intervals cannot be made for mitochondrial genes because they collectively represent a sample size of one gene. Large numbers (dozens or more) of independent loci can produce more accurate and precise parameter estimates than datasets consisting of fewer independent loci. See the studies by Jennings and Edwards (2005), Lee and Edwards (2008), Smith et al. (2013), and Costa et al. (2016) for empirical examples.

3.1.3.1 How Many Independent Loci Exist in Eukaryotic Genomes?

Until recent years, there were few published phylogenomic studies with statistically meaningful numbers of presumably independent and neutral DNA sequence loci (e.g., Chen and Li 2001; Jennings and Edwards 2005; Lee and Edwards 2008). However, thanks to the advent of NGS, researchers are now able to assemble datasets with orders of magnitude more loci for nonmodel organisms than was possible before. For example, NGS-based studies of birds (Faircloth et al. 2012; Jarvis et al. 2014; Prum et al. 2015) obtained hundreds to thousands of presumably independent loci for use in multilocus coalescent analyses in order to infer species trees. Complete genome *in silico* approaches are also generating large datasets consisting of hundreds of independent loci (McCormack et al. 2012; Costa et al. 2016). These enormous datasets show that the genomics era is already making a huge impact on phylogenomics. However, as it is becoming easier and less expensive to obtain large numbers of loci, Costa et al. (2016) asked a new question of importance to phylogenomic studies: *How many genealogically independent loci exist in eukaryotic genomes?*

Let's consider this question using our own genome. Given the vastness of the 3.2-Gb human genome, there would seemingly be an unlimited supply of independent loci. Felsenstein (P. 466, 2004) points out that there may be a million different gene trees that describe the ancestry of the human genome. Let's pause for a moment to think about this figure. While this is a realistic estimate for all gene trees that characterize the evolution of our genome, it is important to realize that this number represents all the different gene trees regardless of whether they are genealogically independent of each other or not. Thus, the term "different" used in this context simply means that the topologies of the trees change as different sites or blocks of linked sites are considered along the lengths of chromosomes. Accordingly, "different gene trees" is not necessarily the same as "genealogically independent gene trees." This conceptual distinction is important. We will soon see that the number of *independent* gene trees in the human genome is a likely a miniscule fraction of a million.

Hudson and Coyne (2002) derived an equation for estimating the number of independent loci or what they termed *independent genealogical units* "IGUs" in a genome, which are defined as the number of loci in a genome whose passage to monophyly is nearly independent of that for all other sampled loci. These workers pointed out that under a neutral model the value of r^2, which is the expected degree of statistical association between two loci, will be $\sim 1/4N_ec$ when $4N_ec$ is large (Ohta and Kimura 1971). Thus, if statistical independence between two loci is achieved when $r^2 = 0.001$, then $4N_ec = 1/0.001 = 1{,}000$. Therefore, the approximate number of IGUs is expected to be:

$$\text{IGUs} = \frac{4N_ec}{1{,}000} \tag{3.1}$$

where N_e is the effective population size and c is the per generation recombination rate between two loci (Hudson and Coyne 2002; Costa et al. 2016 provided this equation in the generic format shown here). Using this equation, Hudson and Coyne (2002) estimated that the common fruit fly (*D. melanogaster*) genome has approximately 11,500 IGUs. How many IGUs are in our genome? Costa et al. (2016) performed the following calculations in order to estimate the number of IGUs in the human genome. The linkage map for humans is 3,614 centimorgans or "cM" (Kong et al. 2002) and estimates for the effective population size range from 7,500 (Tenesa et al. 2007) to 10,000 (Takahata 1993). Using this information and Equation 3.1 the estimated minimum and maximum number of IGUs in the human genome are

$$\text{IGUs}_{\text{min}} = \frac{(4 \times 7{,}500 \times 3{,}614 \text{ cM} \times 0.01 \text{ cross-overs per generation per cM})}{1{,}000}$$

$$\text{IGUs}_{\text{min}} = 1{,}084$$

and,

$$\text{IGUs}_{\text{maximum}} = \frac{(4 \times 10{,}000 \times 3{,}614 \text{ cM} \times 0.01 \text{ cross overs per generation per cM})}{1{,}000}$$

$$\text{IGUs}_{\text{maximum}} = 1{,}446$$

These results suggest that the human genome only contains ~1,000–1,400 independent loci, which is perhaps shocking on two levels. First, it implies that the number of independent gene trees comprises a tiny fraction (~1/1,000) of the total number of gene trees that characterize the ancestry of the human genome. Second, the number of IGUs in the genome of *D. melanogaster* is ten times higher than the number estimated for the human genome, even though the genome size of the former is 18 times smaller than the genome of the later (Brown 2007). However, the difference in the numbers of IGUs between fruit fly and human genomes is explained by the far larger effective population size of *D. melanogaster* ($N_e = 10^6$; Kreitman 1983).

To gain additional perspective on the numbers of IGUs per genome we will now examine several other animal species for which adequate data exist. Table 3.1 shows the relevant population size and genome data for several animal lineages including fruit fly, human, hominoid, zebra finch, grass finch, and tiger salamander. The numbers of IGUs vary substantially largely depending on the assumed N_e values: for tiger salamanders, the number of IGUs is only ~210 if a low N_e is assumed, whereas this number jumps to 21,000 when a larger population size is used to calculate the IGU number (Table 3.1). Species with extremely large N_e such as the zebra finch may have tens of thousands of IGUs as suggested by these calculations. The long-term average N_e for humans is much smaller than the long-term average N_e for the hominoids, which explains why the numbers of IGUs for the former is an order of magnitude lower than the latter (Costa et al. 2016; Table 3.1).

While some of the IGU estimates per genome in Table 3.1 are quite large, it is important to realize that these estimates are for *all* IGUs. This means that IGUs represented by multiple copy and nonneutral loci are included as well (Costa et al. 2016). Because multilocus coalescent-based analyses require that all sampled loci meet the assumptions of being lone orthologous copies, neutral, and independent, the number of *single-copy and neutral IGUs* will likely be far smaller than the maximum number of IGUs (Costa et al. 2016). In other words, when the IGUs that are part of multicopy gene families and are under selection or linked to sites that are under selection are excluded from consideration, the number of single-copy and neutral IGUs will likely be far lower than these maximum numbers. When Costa et al. (2016) used bioinformatics software to find the maximum number of single-copy and presumably neutral IGUs in the hominoid genome, they found only 292 such loci, which is about ~2% of the estimated maximum number of IGUs in the genome (Table 3.1).

These preliminary findings suggest that the maximum number of single-copy and neutral IGUs in animal genomes must be in the hundreds to low thousands and not in the tens of thousands, which raises an important implication. Enabled by NGS-related technologies, researchers are now able to mine genomes for hundreds to thousands of DNA sequence loci. Thus, if a phylogenomics study involves the use of multilocus coalescent methods, then the maximum number of single-copy and neutral IGUs should not be exceeded otherwise the researcher will falsely inflate the number of loci and thereby commit an error known as pseudoreplication (Costa et al. 2016). Among other problems, pseudoreplicating loci will likely result in the researcher obtaining confidence intervals around historical demographic parameter estimates that are incorrect.

3.1.3.2 Criteria for Delimiting Loci with Independent Gene Trees

Because NGS is allowing researchers to harvest unprecedented numbers of loci from genomes—even numbering into the thousands, a need exists for methods that can identify loci that meet the independent loci assumption. In practice, researchers have used two different criteria to delimit independent loci in phylogenomic studies. One criterion, first used by O'Neill et al. (2013), identifies independent loci if they undergo independent assortment, whereas a second criterion developed more recently by Costa et al. (2016) is based on the decoupling of two loci due to long-term effective recombination (Jennings 2016). How do these criteria differ from each other and what are their efficacies?

TABLE 3.1
Estimated numbers of IGUs in various animal genomes

	Fruit fly	Human (low–high)	Hominoid	Zebra finch (low–high)	Grass finch (low–high)	Tiger salamander (low–high)
Population size (N_e)	10^6	7,500–10,000	10^5	1.3×10^6	5.2×10^5	10^3–10^5
Genome map length (cM)	287	3,614	3,614	1,068–1,341	1,068–1,341	5,251
Genome size (bp)	1.8×10^8	3.2×10^9	3.2×10^9	1.2×10^9	1.2×10^9	3.5×10^8
Number of IGUs per genome	11,500	1,100–1,400	14,000	56,000–70,000	25,000–31,000	210–21,000
Number of bp per IGU	15,700	2.3×10^6–2.9×10^6	230,000	17,000–21,000	39,000–48,000	1.7×10^4–1.7×10^6

SOURCE: Data and results obtained: Fruit fly (Hudson, R. R. and J. A. Coyne. 2002. *Evolution* 56:1557–1565.); Human and Hominoid (Costa, I. R. et al. 2016. *Genome Res* 26:1257–1267.); and Tiger salamander (Jennings, W. B. 2016. bioRxiv doi: http://dx.doi.org/10.1101/066332). Data for Zebra and Grass finches were obtained from the following sources: Zebra finch population size (Balakrishnan, C. N. and S. V. Edwards. 2009. *Genetics* 181:645–660.); Zebra finch map lengths (Stapley, J. et al. 2008. *Genetics* 179:651–667; Backström, N. et al. 2010. *Genome Res* 20:485–495.); Zebra finch genome size (Warren, W. C. et al. 2010. *Nature* 464:757–762.); and Grass finch ancestral population size (Jennings, W. B. and S. V. Edwards. 2005. *Evolution* 59:2033–2047.).

NOTE: The term "Hominoid" is used to describe a representative hominoid genome for purposes of estimating the number of IGUs among extant hominoid species (see Costa et al. 2016). The genome size and map length of the grass finch genome is assumed to be the same as for the zebra finch.

If we only focused on loci found on different chromosomes, then these criteria are equivalent because such loci will necessarily have independent gene trees (Wakeley 2009). However, when we consider loci found on the same chromosomes, this becomes an issue of importance for us as we will see below. Though both criteria can identify loci with independent gene trees, one of these criteria is far too stringent—its use can drastically limit the number of loci used in a study.

Let's first examine the methodology developed by Costa et al. (2016). Recall from Equation 3.1 that N_e strongly influences the total number of IGUs in a genome. Thus, if one assumes that N_e and recombination rate are constant, then, in practice, a physical distance in bp (or kb) might be used to delineate IGUs in studies. Because such a minimum physical distance or *distance threshold* between IGUs found on the same chromosome incorporates N_e and recombination rate, it is a direct means for delimiting loci with putatively independent genealogies. For example, from Table 3.1 we see that *D. melanogaster* has an estimated 11,500 IGUs in its genome. Given the 180 Mb genome of *D. melanogaster*, we would therefore expect to see, on average, one IGU every ~15,700 bp along its chromosomes. Thus, sampling a locus every 15,700 bp will yield a dataset consisting of the maximum number of IGUs.

Now let's consider a criterion used to delineate independently assorting loci. From classical genetics, we know that loci separated by 50 cM are effectively unlinked from each other; that is, they are expected to undergo independent assortment the same as if they were on different chromosomes (Hartl and Jones 2006). Applying this criterion to *D. melanogaster*, we see from Table 3.1 that the genome for this species has a map length of 287 cM and thus we could only expect to obtain a maximum of six or seven IGUs. The fruit fly genome has seven chromosomes and so this works out to be ~1 IGU per chromosome! Comparisons between these two criteria using any of the species in Table 3.1 produces comparable results—in all cases the estimated number of IGUs using Equation 3.1 is far higher than when the independent assortment (i.e., 50 cM threshold) is used. Thus, while each criterion does its job in identifying IGUs, the independent assortment criterion is much too stringent and would therefore unfairly restrict the number of IGUs that could be used in a multilocus coalescent-based analysis (Jennings 2016). This assessment regarding the conservative nature of an independent assortment criterion to identify genealogically independent loci agrees with Felsenstein (P. 484, 2004), who noted that the distance between two loci with totally different trees is actually very short, even $\ll$30 cM. In light of these results, it is recommended that researchers not use an "independent assortment" criterion to delimit independent loci in phylogenomic studies (Jennings 2016).

3.1.4 Assumption 4: No Historical Recombination within Loci

Another common assumption in phylogenomic studies holds that, for each locus, there has been *no* recombination within any of the sampled DNA sequences since the time of their most recent common ancestor. This assumption is typically required by multilocus coalescent analyses in phylogeographic (Yang 2002; Rannala and Yang 2003; Edwards et al. 2005; Jennings and Edwards 2005) and species tree studies (Edwards 2009; Lanier and Knowles 2012). This assumption is important because these types of studies consider each gene tree as an independent replicate and thus incorrectly reconstructed gene trees due to past intralocus recombination event(s) is expected to adversely impact these types of studies (Hare 2001; Edwards 2009; Lanier and Knowles 2012). Thus, a locus that has experienced no past recombination since the most recent common ancestor of the sampled sequences is represented by one genealogical history (Wakeley and Hey 1997). For example, as we saw in Chapter 2, all sites in the mitochondrial genome are effectively linked together such that all sites represent one super-locus or super-gene. However, as we will see in Section 3.1.4.1, recombination via crossing over will alter the properties of DNA sequence loci such that multiple distinct gene trees are needed to account for the histories of different sites in a particular recombined "locus."

3.1.4.1 Intralocus Recombination and Gene Trees

It only takes one recombination event within a DNA sequence locus to result in a situation in which two topologically different gene trees are needed to account for the genealogical histories of all sites on the same locus. This is important because many types of phylogenomic analyses assume a one-to-one relationship between loci and gene trees (i.e., one gene tree represents each

locus defined by the researcher). To illustrate this phenomenon, we will borrow the example by Felsenstein (2004), which shows what happens to a hypothetical 204 base long locus after a one intralocus recombination event occurs (Figure 3.3). If the breakpoint due to a recombination event occurred between sites 138 and 139, then *two* gene trees having slightly different topologies are needed to show the genealogical histories of all sites in the locus: one tree is needed to show the history of sites 1–138, while a second similar, but topologically different, tree is needed to account for the history of sites 139–204 (Figure 3.3). Thus, in the aforementioned example the locus consists of two *nonrecombined blocks* of sites: the first block represents sites 1–138 and the second block represents sites 138–204 linked together. In this case, if a researcher unknowingly attempts to infer a gene tree using all the sites from this locus, then the resulting tree can, at best, only reflect one of the two possible tree topologies, which would not capture the complete genealogical histories of all sites.

3.1.4.2 What Is the Optimal Locus Length?

When designing new loci (Chapter 8) the researcher often has to choose the approximate length, in base pairs, of each locus. Given that most loci in nuclear genomes are subject to recombination via crossing over, an important question to consider is *What is the optimal locus length?* Unfortunately, there is no simple answer to this question because a tradeoff exists. On the one hand, longer sequences (all else equal) will tend to produce more resolved and robust gene tree reconstructions because larger numbers of variable sites will contribute to the gene tree estimate. However, a longer locus will also increase the chances of including recombined sites (Felsenstein 2004), which would render such a locus in violation of the no intralocus recombination assumption. While shorter loci will have fewer or no recombination break points within them, these shorter loci will also tend to have fewer informative sites and therefore they will be more susceptible to phylogenetic reconstruction errors. Historically, many of the developed nDNA loci have been less than 1 kb long, an upper bound mostly imposed by the costs and limits of Sanger sequencing. However, new loci development methods are available that enable researchers to design loci of far longer lengths if desired (Chapter 8). Thus, if you are designing your own DNA sequence loci, then how long should they be? Or, if you are using loci developed by others, should you worry that your locus contains more than one nonrecombined block of sites?

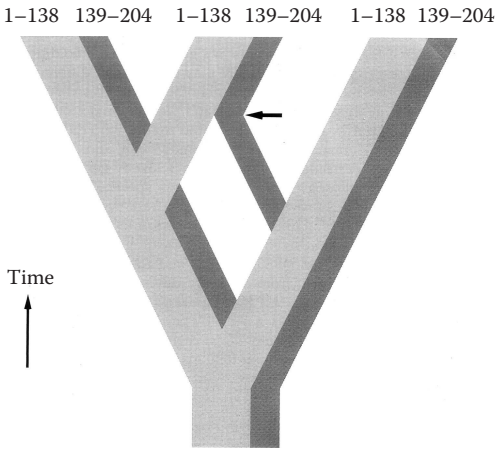

Figure 3.3. Effect of a single recombination event in a locus that is 204 sites long (Felsenstein 2004, fig. 26.8). The legend reads, "A coalescent tree with a single recombination event. The recombination is between sites 138 and 139 at the point indicated by the horizontal arrow. The result is a pair of related coalescent trees." (Reprinted from Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland: Sinauer. With permission.)

Felsenstein (2004) reviewed the theory for understanding the evolutionary factors that determine the physical distance (in base pairs) along a chromosome that is comprised of sites with the same gene tree before a recombination break point is encountered between two adjacent sites. As we saw earlier, the human recombination map length is 3,614 cM (Table 3.1), which means, on average, about 36 cross-overs (or "Morgans") occur in each human genome. Given that the human genome consists of 3.2×10^9 bases (Table 3.1), the expected number of bases between each recombination site is expected to be around 10^8 bases (Robertson and Hill 1983; Felsenstein 2004). However, the true frequency of recombination break points must be evaluated in light of the demographic history of the sample of individual genomes under consideration (Robertson and Hill 1983; Felsenstein 2004). If we consider the genomic sequences of two individuals sampled randomly from the population, then the average number of recombination events that

occurred between them since their last common ancestor is equal to $4N_e r$, where r is the recombination fraction between two points in the genome (Felsenstein 2004). In order to find the long-term average number of bases separating recombination break points (i.e., average nonrecombined block length), we must first set $4N_e r = 1$. The parameter r can then be estimated if a value for N_e is supplied in this formula. Felsenstein (2004) calculated this break-point interval (hereafter "BPI") using the human genome: setting N_e to 10,000 (Takahata 1993), we see that $r = 1/4N_e = 1/40,000 = 2.5 \times 10^{-5}$. This estimate for the long-term average recombination frequency is multiplied by the average number of bases per cross-over per genome (10^8 bases/cross-over), which give us an average nonrecombined block length or BPI of 2,500 bases (Felsenstein 2004). This suggests that, on average, one can expect 2,500 bases to separate each recombination break-point in the human genome (Felsenstein 2004). However, this BPI depends heavily on population size and thus if the human population size is increased an order of magnitude (100,000), then the average interval length is expected to be only 250 bases long (Felsenstein 2004), which is comparable to the original estimate of ~120 bases by Robertson and Hill (1983) who used an estimate of $N_e = 200,000$ in their calculation.

What are the BPIs in the genomes of other species? We can calculate the expected BPI in other species with the following equation based on the work of Robertson and Hill (1983) and Felsenstein (2004):

$$\text{BPI} = \frac{(1/4N_e) \times (\text{genome bp})}{((0.01\,\text{cross-overs/cM}) \times \text{genome cM})} \quad (3.2)$$

Using data for population size, genome size, and genome map length for the zebra finch in Table 3.1 and Equation 3.2 we can calculate the expected BPI in the genome of this species:

$$\begin{aligned}\text{BPI} &= (1/(4 \times (1.3 \times 10^6))) \times (1.2 \times 10^9)/ \\ &\quad (0.01 \times 1{,}341) \\ &= 17 \text{ bp}\end{aligned}$$

Zebra finches have very large population sizes and thus their average nonrecombined block has an expected interval length of only 17 bases long, which is comparable to the estimated ~16 base interval in *D. melanogaster* (calculation not shown). Australian grass finches likely have a similar genome size and recombination map to the zebra finch, but they have smaller population sizes (Table 3.1) and thus their BPI is expected to be larger at ~54 bases (calculation not shown). If we assume that most animals have population sizes between humans and fruit flies, then the average length of a nonrecombined block of sites is expected to range from a maximum of ~2,500 bases down to a minimum of ~16 bases. This is an alarming result because gene trees are routinely inferred from individual loci that span hundreds or more bases. Lanier and Knowles (2012) suggested that the no recombination assumption is commonly violated in species tree studies and the calculations above shows us why this must be so. However, as Felsenstein (2004) pointed out, recombination rates vary across genomes with some regions having hotspots and others coldspots (e.g., McVean et al. 2004). We can use this fact to our advantage in constructing phylogenomic datasets consisting of nonrecombined loci, at least for recently diverged species.

One remedy is to identify presumably nonrecombined blocks of sites within a sequence dataset and then retain only the largest nonrecombined block for downstream gene tree analyses (Jennings and Edwards 2005). Hudson and Kaplan (1985) developed a method called the "four-gamete test," which can identify the *minimum* number of past recombination events within a DNA sequence locus. To see how this method works, we will examine a simple example that is illustrated in Figure 3.4. In this example, we have a locus that is 52 bases long and is represented by four DNA sequence alleles. In reality, loci to be tested will usually have hundreds or more contiguous sites. Nonetheless, for illustrative purposes we can still evaluate this short locus using the four-gamete test. The first step is to identify all the segregating (i.e., polymorphic or variable) sites, which in our locus includes sites 4, 20, and 49 (Figure 3.4a). In the next step, we perform pairwise comparisons between these polymorphic sites and then count the number of "gametes" or unique 2-base combinations: doing this shows us that sites 4 and 20, considered together, yield four unique base combinations, while the other pair of sites (20 and 49) only have three unique gametes each (Figure 3.4b). In the final step, we count one recombination event for every pair of sites showing all four possible gametes and then sum

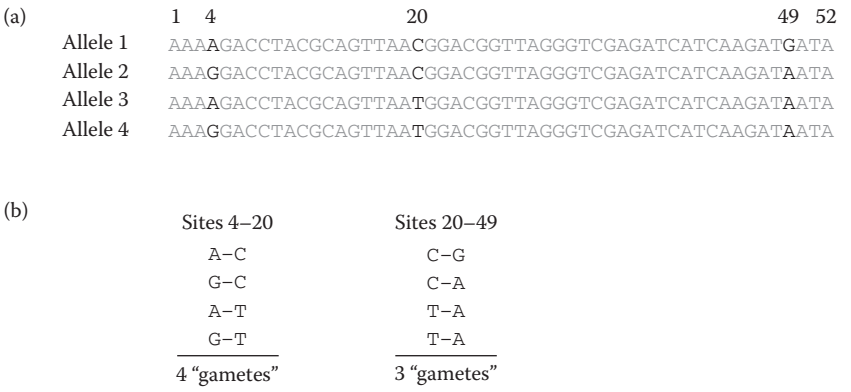

Figure 3.4. The four-gamete test of Hudson and Kaplan (1985). (a) The test is applied to a minimum of four orthologous DNA sequences. In this example, the sequences are 52 base pairs (sites) long with three segregating sites (i.e., 4, 20, and 49). The test only considers the segregating sites (black letters) and not the monomorphic sites (gray) in a sample. (b) Two pair-wise comparisons between consecutive segregating sites are needed to complete a test for the minimum number of historical recombination events in this sample (R_m). The comparison between sites 4 and 20 reveal four different "gametes," which is indicative of a past recombination event between these sites, whereas the comparison involving sites 20 and 49 show only three unique gamete types, which suggests no recombination occurred between these sites. Thus, the estimated minimum number of historical recombination events in these sites is $R_m = 1$.

them to obtain a value for the R_m or the minimum number of past recombination events. Our results yield an $R_m = 1$, which means that we have identified one historical recombination event within our locus. This result tells us that in the past there was at least one recombination event that occurred somewhere between sites 4 and 20 in our locus. We cannot know exactly where the breakpoint was because these intervening sites are all invariable. However, we can still use this information to modify our locus so that it conforms to the no intralocus recombination assumption. The procedure is simply to truncate the locus down to the largest nonrecombined block of sites, which in our example is defined by sites 20–52. The other sites from the smaller block, which includes the recombined region, are discarded. Although this test is easy enough to perform by hand on small datasets (e.g., with the minimum four sequences or a low number of polymorphic sites), it is better to search sequences for evidence of historical recombination events using a computer program such as *DNAsp* (Rozas et al. 2003), *IMgc* (Woerner et al. 2007), or *RDP3* (Martin et al. 2010).

There are several reasons why this statistic represents an estimate of the *minimum* number of past recombination events. First, back mutation can convert a polymorphic site to an invariable site and thus possibly render a past recombination event invisible to the test. Secondly, a double-recombination event in the same intervening region between two consecutive polymorphic sites can also "erase" the recombination history in this genomic segment. Lastly, although the minimum number of sequences required for the four-gamete test is four, it is always possible that using a larger number of sequences will reveal new polymorphic sites and hence this could conceivably lead to the discovery of additional past recombination events. Thus, the four-gamete test will most likely underestimate the true number of past recombination events. However, as Lanier and Knowles (2012) point out, this is not expected to be a problem because not all recombination events alter the topology of a gene tree. Therefore, this test should only be used for purposes of determining whether a given locus dataset shows evidence of past recombination. If the test is applied to a dataset and *Rm* is estimated to be zero, then the original locus can be assumed to meet no intralocus recombination assumption, whereas if *Rm* > 0, then the largest nonrecombined block can be identified and retained for further coalescent-based analyses.

An important assumption of the four-gamete test holds that each site has only mutated once and therefore follows the "infinite sites" model of sequence evolution. According to the infinite sites model, each site can only undergo a single substitution. However, if any sites in a locus have

undergone multiple substitutions, then the four-gamete test may yield false positive results and thus overestimate the number of historical recombination events in a locus. False positive results can also be caused by DNA sequencing errors, haplotype phasing errors, and sequence alignment errors (Martin et al. 2010).

One problem with this approach is that by truncating loci down to the largest presumably nonrecombined block, you solve one problem (i.e., meet the no recombination assumption) but create another. Truncating loci means eliminating variable sites, which will elevate the incidence of phylogenetic reconstruction errors (Lanier and Knowles 2012). In effect, eliminating one type of gene tree reconstruction error simultaneously causes another type of gene tree error. Multilocus coalescent methods used in phylogeographic and species tree studies often rely mainly on the individual gene tree topologies and thus topological errors whether they are due to recombination or lack of phylogenetic signal in the sequences may adversely affect results.

One approach that can be used to solve both problems takes advantage of the heterogeneity of recombination rates across genomes. Although an average expected nonrecombined block may be short (e.g., 50 bp), it will likely be possible to find longer blocks, which can be used in phylogenomic analyses. Thus, an initial dataset can be collected, which consists of loci that are fairly long (e.g., >2 kb). Recombination tests are then applied to find the longest nonrecombined blocks in each locus. The longest nonrecombined block is then retained from each locus (one block per locus) for downstream analyses. This approach would help reduce the adverse effects of recombination and gene tree reconstruction errors. It should be noted, however, that this methodology will only be useful for datasets consisting of recently diverged species (i.e., sequences that conform to the infinite sites model). However, given that most multilocus coalescent-based studies involve closely related species, this should not be a problem. Unfortunately, it may not be possible to identify recombined sites in loci consisting of more anciently diverged sequences because multiple hits or alignment errors due to indels would preclude the use of recombination tests, which assume an infinite sites model. On a more encouraging note, the simulation study of Lanier and Knowles (2012) suggests that robust species tree inferences can be obtained even if the intralocus assumption is violated. Future studies should perform sensitivity analyses whereby multilocus coalescent-based analyses are conducted on recombination-corrected and uncorrected datasets to determine how sensitive the resulting phylogeographic parameters or species trees are to the no intralocus recombination assumption.

3.1.5 Assumption 5: Loci Evolved Like a Molecular Clock

Zuckerlandt and Pauling (1965) noticed that substitution rates for amino acids in different hemoglobin chains for human, horse, and cattle appeared to be constant. This observation of lineage rate constancy of substitutions prompted these authors to propose the *molecular evolutionary clock hypothesis*. Although this was originally proposed for protein data, the molecular clock hypothesis has since been extensively studied in DNA sequences. The clock-like manner by which substitutions seem to occur in some genomic sites is a useful property for a DNA sequence locus to have. The molecular clock is often used to estimate the timing of historical events, which can be gene divergences or population and species divergences (Swofford et al. 1996; Arbogast et al. 2002; Yang 2006). The molecular clock is also a standard assumption of coalescent-based gene tree analyses (Felsenstein 2004) and it can even be used to root gene trees (Huelsenbeck et al. 2002; Jennings et al. 2003; Felsenstein 2004; Yang 2006). If the clock is used to date a historical divergence event, then the mutation rate must be calibrated, which means that the units of substitutions must be converted into time units. These calibrations are obtained from time-dated fossils.

The so-called "strict" molecular clock assumption states that the mutation rate (μ) is constant among all sequences in a sample (i.e., among lineages). Although this condition is expected to be observed among closely related species owing to their similar physiologies, DNA repair, and generation times, the clock may gradually break down for more divergent species (Felsenstein 2004; Yang 2006). In cases where molecular data do not meet the strict clock assumption, investigators have used other clock-like methods such as *relaxed molecular clocks* (Felsenstein 2004) and *local clock* approaches (Thorne and Kishino 2005; Yang 2006).

3.1.6 Assumption 6: Loci Are Free of Ascertainment Bias

Molecular systematists and population geneticists have long known that some loci are more variable than other loci. Moreover, because variable loci contain "more information" than less variable loci, it may be tempting to choose the most variable loci for a study while discarding less variable ones. However, this act of "cherry picking" the most variable loci can lead to a form of bias known as *ascertainment bias* (Kuhner et al. 2000; Nielsen 2000; Wakeley et al. 2001; Brumfield et al. 2003; Felsenstein 2004; Nielsen et al. 2004; Rosenblum and Novembre 2007). If loci or sites are chosen nonrandomly, then it is important to try and correct the ascertainment biases (e.g., Kuhner et al. 2000; Nielsen 2000; Wakeley et al. 2001; Felsenstein 2004).

3.2 DNA SEQUENCE LOCI: TERMINOLOGY AND TYPES

Now that we reviewed the composition of genomes, various aspects of molecular evolution, and the common assumptions for DNA sequence loci used in phylogenomic studies, we are finally ready to review the types of loci commonly used in studies. As was already mentioned, studies focusing on the evolution of genes or genomes will know right away which locus or loci to use, whereas studies directed at reconstructing the tree of life must choose from among the various loci types. Accordingly, much of the following discussion will be mainly relevant for those interested in tree of life studies. Later in this book, in Chapters 7 and 8, we will see how to acquire these loci. Before we start this discussion, however, we must address an important issue relevant to everyone conducting phylogenomic studies, which concerns the inconsistent use of some commonly used genetics terms.

3.2.1 On Genes, Alleles, and Related Terms

The terms *gene*, *locus*, *allele*, *gene copy*, and *haplotype* are not always used in a consistent manner, which is unfortunate because misusing these terms can lead to confusion when trying to understand phylogenomic concepts and studies. Gillespie (P. 6–10, 2004) recognized this problem and brought much needed clarity. However, this issue is important enough that it deserves to be revisited here.

First, the meaning of the word "gene" has varied depending on the background of the worker. For example, molecular biologists use the traditional definition of gene and thus only regard functional segments of DNA as being genes such as RNA-coding and protein-coding sequences (e.g., Brown 2007; Watson et al. 2014). Population geneticists, on the other hand, tend to use this term more broadly with some workers relying on the function-based definition because it relates to some organismal trait, while others consider *any* genomic segment—even one base pair—regardless of whether or not it has a known function as a gene (e.g., Felsenstein 2004; Wakeley 2009). As Gillespie (2004) pointed out, the term "locus" refers to a specific genomic location where a segment of DNA is located. Thus, to the molecular biologist and some population geneticists, a gene is a functional segment of DNA and locus refers to its genomic location. In contrast, other population geneticists may refer to any segment of DNA—regardless whether it is functional or not—as a gene or locus. Owing to the confusion surrounding the various uses of the word gene, Gillespie (2004) suggested that this term has lost its usefulness in population genetics studies. However, given the growing popularity of phylogenomic studies that use the term "gene trees," in this book we will retain the term gene and use it interchangeably with locus.

The term "allele" has long been a fundamental unit in genetics, but what exactly is an allele? If we define a gene (or locus) as a particular segment of genomic DNA, then the allele is the actual physical manifestation (DNA sequence) of the gene found in a particular individual. Another term that is sometimes used *in lieu* of allele, particularly in population genetics, is *gene copy* (e.g., Felsenstein 2004). A third term that is used synonymously with allele and gene copy is *haplotype* owing to the haploid nature of an allele or gene copy. The term haplotype has been used often in phylogeographic studies to describe alleles perhaps because of the predominant use of mitochondrial DNA during the first two decades of this discipline (e.g., Avise 2000). As you will recall from Chapter 2, any DNA sequence obtained from a mitochondrial genome is a haplotype because there is only one copy of the mitochondrial genome inherited from a parent (most often the mother). But if we look at a human autosomal gene, we cannot forget that there will be *two copies* of that gene in each

individual (i.e., one from each parent). Although restating this fundamental of genetics may seem unnecessary, we will see elsewhere in this book some cases in which it is not uncommon for researchers to overlook this idea thereby leading to some confusion. Throughout this book we will consider the terms allele, gene copy, and haplotype as having identical meanings and it does not matter whether an allele is a 500 bp sequence or a single base.

Unfortunately, the term gene (or locus) has also been confused with allele (or gene copy or haplotype). For example from medicine, if a doctor says that a person is "gene positive" for Huntington's disease (HD), what the doctor is really saying is that the person has inherited the *allele* (or gene copy or haplotype) that gives rise to the disease (HD is an autosomal dominant disease). Most people inherit two normal (nondisease) alleles from their parents, but those that are so-called gene positive inherit a copy of the defective allele. In an example from population genetics, it is not uncommon to see an author talking about genes in a phylogenetic study but in reality the author is describing gene *copies* (or alleles or haplotypes) for a gene. Needless to say, this can become confusing if a study includes numerous genes (or loci) each of which is represented by multiple haplotype sequences. Thus, it is essential for phylogenomics researchers to use these terms in a consistent manner.

3.2.2 Commonly Used DNA Sequence Loci in Phylogenomic Studies

3.2.2.1 Mitochondrial DNA Loci

Ever since the late 1970s and continuing to the present day, mitochondrial DNA loci have been the most extensively used genetic markers in molecular phylogenetic studies. Despite the rapid growth of genomics over the past decade, mitochondrial loci are still tremendously popular evolutionary markers owing to their many desirable properties. As we saw in Chapter 2, these properties include their haploid nature, negligible recombination, high substitution rate relative to nuclear sites, and low effective population size compared to most nuclear loci. Regarding this last property, the relatively smaller effective population sizes of mitochondrial DNA implies that, on a locus per locus basis, mitochondrial loci will better track the species tree than will autosomal loci (Moore 1995; Zink and Barrowclough 2008). Note, however, that despite claims of mitochondrial DNA being "selectively neutral," the earlier discussion in this chapter about the indirect effects of natural selection (e.g., background selection) show why this is very unlikely to be true (Wilson et al. 1985; Ballard and Rand 2005; Galtier et al. 2009). Another contributing factor to the popularity of mitochondrial DNA, especially during the early years of molecular systematics, was the introduction of the first universal PCR primers by Kocher et al. (1989), which allowed researchers to obtain mitochondrial sequence data from a wide variety of metazoan species. Not only have mitochondrial loci played a huge role in molecular systematics, but they also were largely responsible for launching phylogeography (Avise 2000) and DNA barcoding (Hebert et al. 2003).

Although nuclear loci are rapidly gaining momentum in phylogenomics studies, it should not be forgotten that mitochondrial DNA will always provide an independent evolutionary perspective of historical events and thus it can complement the perspective offered by nuclear loci (e.g., Prychitko and Moore 1997; Reilly et al. 2012). The following summary of mitochondrial loci is not comprehensive. For reviews and perspectives on the utility of mitochondrial DNA in different types of phylogenomic studies the reader should consult the works by Wilson et al. (1985), Moritz et al. (1987), Moore (1995), Palumbi (1996), Avise (2000), Hebert et al. (2003), Ballard and Rand (2005), Zink and Barrowclough (2008), and Galtier et al. (2009).

Mitochondrial protein-coding loci—Mitochondrial protein-coding loci have been perhaps the most important class of mitochondrial DNA. As we saw in Chapter 2, the existence of codon bias in protein-coding genes means that different codon positions will exhibit different levels of variability. This aspect of their evolution is of practical importance because it can be useful for reconstructing both recent and old divergences in a gene tree. Another nice feature of protein-coding sequences, at least compared to most noncoding sequences, is that multiple sequence alignments with protein-coding sequences are usually simple to perform even among highly diverged sequences. The mitochondrial *cytochrome oxidase I* gene or "COI" has been a particularly important mitochondrial gene due to its central role in the DNA barcoding program (Hebert et al. 2003).

Mitochondrial ribosomal RNA loci—The mitochondrial genomes of animals contain two ribosomal RNA (rRNA) genes (Mindell and Honeycutt 1990; Palumbi 1996). One of the gene sequences codes for the 12S rRNA or "small subunit" while the other codes are for the 16S rRNA or "large subunit" (Palumbi 1996). In vertebrates the 12S rDNA exists as a ~950 bp long sequence, whereas the 16S rDNA is ~1,600 bp long. In contrast to nuclear rDNA sequences, the 12S and 16S sequences in the mitochondrial genome are contiguous (do not contain any spacer DNA; Mindell and Honeycutt 1990). In order to obtain 12S and 16S sequences, researchers typically acquire these gene sequences from ~300 to 800 bp loci, which yield partially overlapping sequences that are later joined into continuous sequences or "contigs" during data analysis. Palumbi (1996) provides lists of widely used 12S and 16S PCR primers—most of which perform well on diverse metazoan species—as well as information pertaining to their physical locations and orientations in the mitochondrial genome.

Both the 12S and 16S rDNA segments in the mitochondrial genome contain stretches of bases that are extremely conserved (i.e., "stem" regions) and stretches that are highly variable (i.e., "loop" regions; Palumbi 1996). Thus, there is great heterogeneity in the evolutionary rates among sites. The among-site variation within these genes makes the process of multiple sequence alignment easy or difficult depending on which stretches of sites are being aligned. The highly conserved blocks are simple to align while other sections are difficult if not impossible. For stretches of sites that have questionable positional homologies, Swofford and Olsen (1990) recommend discarding them because it is safer to remove these sites rather than to infer a gene tree based on spurious alignments.

Despite these minor complications, both the 12S and 16S genes have long been favorite DNA sequence loci of molecular systematists because of their effectiveness in resolving the phylogenetic relationships within eukaryotes particularly among species, genera, and families (e.g., Mindell and Honeycutt 1990; Hillis and Dixon 1991; Palumbi 1996; Darst and Cannatella 2004). The mitochondrial 16S rRNA gene also has special properties, which makes it an attractive candidate for being a DNA barcode locus in amphibians (Vences et al. 2005).

Mitochondrial control region loci—The control region is the most variable part of the mitochondrial genome (Vigilant et al. 1989; Tamura and Nei 1993; Palumbi 1996; Randi 2000). The control region is hypervariable because, unlike the protein-coding regions, it can tolerate both base substitutions and indels of varying lengths yet still remain functional (Vigilant et al. 1989). However, this property of control region loci means they are a poor choice for studies of highly divergent species owing to difficulties with multiple sequence alignments. Also, attempts to obtain high quality control region sequences using the Sanger method can be frustrated by the existence of indels (Chapter 6) that reside either in different heteroplasmic copies of the mitochondrial genome or from artifacts in PCR due to long microsatellite repeats. Despite these drawbacks, control region DNA has been an effective evolutionary locus for studies operating at shallow phylogenetic scales especially those at the intraspecific level (Vigilant et al. 1989, 1991; Baker et al. 1993; Tamura and Nei 1993; Avise 2000; Randi 2000). Control region loci have also been employed as DNA barcode-like sequences in conservation genetics studies (Baker and Palumbi 1994). For further information about the structure and organization of the control region the reader can consult Avise (P. 27, 2000) for an illustration of the entire mammalian control region. Palumbi (1996) provides additional details about control region loci not discussed here.

3.2.2.2 Nuclear DNA Loci

The nuclear genome is a veritable treasure trove for DNA sequence loci yet it has largely remained inaccessible to researchers until recent years. This is because prior to the commencement of the genomics era, the development of novel DNA sequence loci was accomplished using the slow and technically challenging methodology of gene cloning. However, in recent years the availability of genomics resources has empowered researchers with a variety of exciting new approaches to developing phylogenomic loci and acquiring datasets consisting of hundreds or more loci. We will review these approaches in Chapters 7–9.

Nuclear loci enjoy important advantages over mitochondrial loci. Besides the sheer variety and number of loci that can be obtained from the nuclear genome, some nuclear loci have the

desirable properties of being presumably neutral and genealogically independent of other sampled loci. These two properties, which are not found in mitochondrial loci, enable researchers to conduct statistically rigorous multilocus analyses in tree of life studies. However, nuclear loci are not without disadvantages. For example, earlier we saw how intralocus recombination can complicate gene tree inferences. In Chapter 6, we will see how the presence of multiple haplotypes at each diploid (or polyploid) locus can also complicate matters though NGS-based methods for data acquisition are largely immune to this problem. Despite these issues, tree of life studies are now making extensive use of large numbers of nuclear loci and this trend will increase further as access to genome-scale datasets becomes simpler, faster, and less expensive in coming years. The following is a brief introduction to the common types of nuclear loci used in phylogenomic studies. For more extensive and in depth discussions of these loci the reader should see the reviews by Palumbi (1996) and Thomson et al. (2010) as well as the references cited below.

Exon-primed-intron-crossing loci (EPICs)—Motivated by the limitations of mitochondrial DNA, researchers attempted to find alternative loci in the nuclear genome so that additional independent perspectives on the evolutionary history of populations or closely related species could be obtained. Nuclear introns represented one such important new class of DNA sequence locus. Introns are largely evolutionarily unconstrained except for their highly conserved 3′ and 5′ termini, which consist of functionally critical sites (Fedorov et al. 2002). Thus, introns display a high level of variability compared to protein-coding and RNA-coding loci, which makes them ideal for studies of populations or closely related species. With this property of introns in mind, Lessa (1992) developed PCR-based intron loci and illustrated their applicability to the study of pocket gopher (*Thomomys bottae*) populations in California. Moreover, to increase the chances of PCR success across different genetic samples, Lessa designed the two PCR primers in highly conserved exon regions flanking the intron of interest rather than nesting them within the intron itself. Palumbi and Baker (1994), who developed their own universal exon–intron loci for vertebrates, called these loci *Exon-Primed-Intron-Crossing loci* or "EPICs," a label subsequently adopted by researchers (Shaffer and Thomson 2007). Many studies have confirmed that intron loci are highly variable even at the population level (e.g., Lessa 1992; Palumbi and Baker 1994; Prychitko and Moore 1997; Sequeira et al. 2006). EPIC loci have generated high-resolution gene trees that are comparable to mitochondrial loci (Palumbi 1996; Weibel and Moore 2002).

Although EPIC loci have enjoyed success in many studies, they suffer from some problems that do not similarly affect mitochondrial loci. Aside from the multiple haplotype per individual problem already mentioned, EPIC loci are largely free to accumulate multiple base substitutions and indels, which can complicate efforts to align sequences from highly diverged species such as above the genus level (Prychitko and Moore 1997; Sequeira et al. 2006 and references therein). Indels of varying lengths are commonly observed in introns (e.g., Ohresser et al. 1997; Hassan et al. 2002) and the more evolutionarily divergent the species the more difficult it is to make an acceptable alignment. However, for intraspecific studies or interspecific studies involving closely related species, multiple sequence alignments are usually trivial (e.g., Sequeira et al. 2006).

Notwithstanding, EPIC loci still retain a number of advantages over mitochondrial loci. First, because mitochondrial loci collectively represent one independent locus (Wilson et al. 1985), such loci cannot be used to infer all details about the evolutionary history of populations or species (Edwards and Bensch 2009). Instead, many independent nuclear loci are required (e.g., Edwards and Beerli 2000). Despite their difficulties, EPICs can better capture the evolutionary history of populations or closely related species because a large number of independent EPIC loci can be harvested from genomes. Many researchers have suggested that most of the sites within introns are nearly neutral in character (Lessa 1992; Prychitko and Moore 1997; Friesen 2000; Daguin et al. 2001; Fedorov et al. 2002). However, as we saw earlier this chapter (see Figure 3.1), the genetic variability of introns is likely to be reduced due to the effects of indirect natural selection acting at nearby sites. Therefore, EPIC loci may not be appropriate for phylogenomic analyses that require the neutrality assumption.

Many EPIC loci have PCR primers that are "universal" in character, which means they can be used to acquire the same sequences from a wide range of metazoan species (e.g., Chenuil et al. 2010). We

will discuss how universal primers are made in Chapter 8 but for now it suffices to know that such primers greatly simplify the acquisition of DNA sequence datasets. Fully sequenced EPIC loci contain stretches of highly conserved exon sequences that flank the entire intron sequence nested inside (Lessa 1992). If the entire sequence is analyzed, then the mixture of high and low variation sites can allow for robust reconstructions of gene trees at both shallow (e.g., intraspecific) and deeper phylogenetic levels (Li et al. 2010). Alternatively, gene trees could be inferred using only the exon sequences if the intron sequences are not alignable (e.g., Li et al. 2010) or the 3′ and 5′ exon sequences can be trimmed away so that only the intron sequences are used in phylogenetic analyses (e.g., Sequeira et al. 2006). For additional information about EPIC loci see Palumbi (1996) and Friesen (2000).

Anonymous loci—One year after the publication of the first EPIC paper by Lessa (1992), Karl and Avise (1993) proposed a new class of DNA sequence loci called *single-copy nuclear anonymous loci*, which could be applied to the study of populations. These so-called "anonymous loci" were given this name because they were developed from random and unknown genomic locations (Karl and Avise 1993; Jennings and Edwards 2005).

Anonymous loci have several advantages that make them ideal for studies involving populations or closely related species. First, because most anonymous loci reside in intergenic regions, they are thought to be free from the influence of natural selection (Thomson et al. 2010; Costa et al. 2016). Anonymous loci are therefore expected to exhibit higher levels of genetic variation than loci under the direct or indirect effects of selection (see Figure 3.1 and Costa et al. 2016 for empirical evidence). Indeed, among all types of phylogenomic loci currently available, anonymous loci are the only loci that may meet the neutrality assumption (i.e., totally free of the effects of natural selection). Another advantage is that large numbers (e.g., hundreds or more) of single-copy and genealogically independent anonymous loci can be mined from genomes (Chapters 8 and 9; Costa et al. 2016). Finally, because anonymous loci are obtained without regard to their variability, they are free from ascertainment biases.

Like intron loci, anonymous loci have a number of drawbacks such the difficulty in aligning sequences from highly diverged species, susceptibility to the effects of intralocus recombination, and difficulties in haplotype resolution. Despite this, anonymous loci still represent the ideal locus type for phylogenomic studies that employ multi-locus coalescent-based analyses because the gene trees for these loci are expected to best resemble the structure of coalescent trees (see Felsenstein 2004 and Wakeley 2009 for reviews of coalescent theory).

Anonymous loci have been developed in a wide variety of vertebrates including hominoids (Chen and Li 2001; Costa et al. 2016), Australian grass finches (Jennings and Edwards 2005), eastern fence lizards (Rosenblum et al. 2007a,b), turtles (Shaffer and Thomson 2007), Australian fairy wrens (Lee and Edwards 2008), black salamanders (Reilly et al. 2012), sea snakes and geckos (Bertozzi et al. 2012), chorus frogs (Lemmon and Lemmon 2012), Atlantic Rainforest antbirds (Amaral et al. 2012), and Mojave fringe-toed lizards (Gottscho et al. 2014), among many others. As we will see in Chapters 8 and 9, the use of anonymous loci is expected to dramatically increase as more researchers take advantage of new NGS and complete genome methods for obtaining large anonymous loci datasets.

Nuclear protein-coding exon loci (NPCLs)—In contrast to introns and anonymous loci, nuclear exons represent a class of more evolutionarily conserved DNA sequence loci. Nuclear exons are also referred to as *Nuclear protein-coding exon loci* or "NPCLs" to distinguish them from organellar protein-coding loci (Thomson et al. 2010). Because of their conserved nature and general lack of indels (indels, if they occur, must be in multiples of 3-bp to preserve the correct reading frame), NPCLs sequences are usually easy to align even for some highly diverged species. NPCLs usually do not show sufficient variation among sequences at the intraspecific level or among closely related species for producing well-resolved gene trees. Thus, the variation present in NPCLs is at a level best suited to reconstructing the gene trees for highly diverged species (e.g., genus or family). For example, in a molecular phylogenetic study of Australian pygopodid lizards, a gene tree inferred from nuclear *c-mos* gene (Saint et al. 1998) sequences did not resolve all species relationships within each genus but it did recover the same generic clades that had been found in a mitochondrial gene tree and a tree based on morphological characters (Jennings et al. 2003). Thus, NPCLs

will be better choices for phylogenomic studies involving speciation events that occurred tens of millions of years ago or earlier.

Until recently, the number of NPCL loci available to researchers was limited. However, once genomics resources (e.g., annotated full genome sequences) for a variety of organisms became available, researchers began using bioinformatics-based methods to design large numbers of NPCLs that could be applied to a wide array of nonmodel organisms (reviewed in Chapter 8). For example, Li et al. (2007) developed 154 presumably single-copy NPCLs that can be applied to a wide variety of fish species. In a similar study, Portik et al. (2011) generated 104 NPCL loci that can be used to obtain sequence data for squamate reptiles.

Traditionally, NPCL loci have been PCR-amplified from genomic DNA followed by Sanger sequencing. However, in recent years researchers have developed exciting new NGS-enabled *transcriptome*-based methods for acquiring hundreds or thousands of exonic sequences from multiple individuals in a single experiment. Thus, rather than directly PCR-amplifying exons from genomic DNA templates, extracted mRNA transcripts are first converted into a *cDNA library* before being sequenced using NGS methods (Chapter 7). For example, Bi et al. (2012) used this approach to obtain DNA sequence data from over 10,000 exonic loci for *Tamias* chipmunks.

Nuclear ribosomal RNA loci—Nuclear ribosomal RNA genes exhibit a number of characteristics that differentiates them from mitochondrial rRNA genes. First, these genes are members of an "array," which consists of three rDNA coding segments (18S, 5.8S, and 28S), two internal transcribed spacer elements (ITS-1 and ITS-2), an external transcribed spacer (ETS), and a non-transcribed spacer (NTS) flanking each end of the array. The structure of this array, which is highly conserved in eukaryotes, consists of the following ordering of elements starting at the 3′ end of the rDNA coding strand: 3′—NTS, ETS, 18S, ITS-1, 5.8S, ITS-2, 28S, NTS—5′ (Mindell and Honeycutt 1990; Hillis and Dixon 1991). In contrast to mitochondrial rRNA genes, which occur as single copies in the mitochondrial genome, nuclear rDNA arrays have been tandemly duplicated hundreds or thousands of times depending on species (Mindell and Honeycutt 1990; Palumbi 1996). An NTS element separates each array copy within a cluster of arrays on a chromosome (Mindell and Honeycutt 1990; Palumbi 1996). Because these array copies have been subjected to the homogenizing effects of concerted evolution, DNA sequences obtained from any array copy can still be effective at resolving the phylogenetic relationships among organismal lineages (Hillis and Dixon 1991; Moritz and Hillis 1996). Another difference between nuclear and mitochondrial rDNA genes is that the former are more highly conserved than the latter and thus they complement each other in terms of the phylogenetic time depths that each can resolve; specifically, nuclear rDNAs are best for pre-Cenozoic divergences (>65 million years ago) while mitochondrial rDNAs are better for studying the evolutionary diversification of groups within the Cenozoic (Hillis and Davis 1986; Mindell and Honeycutt 1990; Hillis and Dixon 1991). Some regions of rDNA genes are so highly conserved that they have revealed the basic domains of life (Woese and Fox 1977; Woese et al. 1990) and resolved the evolutionary relationships among phyla (Hillis and Dixon 1991). Another desirable property of nuclear rDNA arrays is that the different elements exhibit variable conservation of sites and thus researchers can target particular elements for sequencing depending on the presumed levels of evolutionary divergence among the study organisms in a given study (Hillis and Dixon 1991). The rDNA genes exhibit the highest levels of conservation followed by the transcribed spacer elements and then the NTS elements (Hillis and Dixon 1991; Palumbi 1996). Moreover, there is also substantial among site variability within and among the rDNA genes (Hillis and Dixon 1991). Lists of PCR primers and guidance for sequencing entire rRNA arrays can be found in Hillis and Dixon (1991) and Palumbi (1996). For reviews of rDNA structure, evolution, and phylogenetic utility the reader should consult the reviews by Mindell and Honeycutt (1990) and Hillis and Dixon (1991).

Anchored loci—The development of *ultraconserved elements-loci* or "UCE-loci" (Faircloth et al. 2012) and *anchored enrichment-loci* or "AE" loci (Lemmon et al. 2012; Lemmon and Lemmon 2013) represents two exciting and important phylogenomic innovations that arose in recent years. Although the methods used to design UCE-anchored and AE loci involve different approaches (see Faircloth et al. 2012 and Lemmon et al. 2012 for descriptions of protocols), these loci share many evolutionary properties and occupy similar phylogenomic loci

niches. Thus, for simplicity we will generically refer to them as *anchored loci*. The significance of the term "anchor" will be made clearer below.

Unlike the previously discussed loci, which are commonly sequenced using the traditional PCR-Sanger sequencing routine, anchored loci are sequenced using NGS-based methods to which they are perfectly suited. Individual sets of UCE-anchored and AE loci consist of hundreds or thousands of single-copy and presumably independent DNA sequence loci found throughout the genomes of various eukaryotic groups. Studies using anchored loci have typically generated datasets consisting of sequences for hundreds to thousands of loci from dozens (e.g., Jarvis et al. 2014) to hundreds (e.g., Prum et al. 2015) of individuals. Phylogenomic studies based on these enormous datasets are resolving species trees at both shallow (i.e., from hundreds of thousands to around ten million years ago) and deep (i.e., tens to hundreds of millions of years ago) levels of evolutionary divergence (e.g., Faircloth et al. 2012; Crawford et al. 2012; Faircloth et al. 2013; McCormack et al. 2013; Jarvis et al. 2014; Prum et al. 2015). Thus, anchored loci studies are already dramatically increasing and refining our knowledge of the tree of life. However, what exactly are UCEs and AEs? We will first consider UCEs and UCE-like elements because they have attracted considerable interest of genome researchers in recent years.

The term "ultraconserved element" was originally coined by Bejerano et al. (2004) to describe the hundreds of mostly noncoding genomic elements common to the human, mouse, and rat genomes, which are perfectly conserved (100% identical) for at least 200 bp. Nearly all of these UCEs also exist in the chicken and dog genomes, albeit they are slightly less conserved than the human-rodent UCEs, and some are found in fish (Bejerano et al. 2004). Moreover, thousands more UCEs that are perfectly conserved for more than 100 bp are found throughout mammals (Bejerano et al. 2004). Human UCEs are most often found to be overlapping exons of RNA processing genes or they exist in intergenic regions adjacent to regulatory and developmental genes (Bejerano et al. 2004). Another notable feature of human-rodent UCEs is that the vast majority of them exist as single-copy elements (Bejerano et al. 2004; Derti et al. 2006; Faircloth et al. 2012). Several other classes of highly conserved genomic elements displaying levels of sequence conservation below that of UCEs have since been discovered in comparative genome studies involving various eukaryotic species.

Siepel et al. (2005) defined a class of conserved genomic elements called "highly conserved elements" or "HCEs," which exhibit slightly less sequence conservation and tend to be longer than human-rodent UCEs. These workers performed a comparative genome analysis in an effort to locate all HCEs within the genomes from representative groups of eukaryotes (i.e., vertebrates, insects, worms, and yeasts). The resulting group-specific sets of HCEs, which they found, ranged in size from five to thousands of bp with an average size of 100–120 bp (Siepel et al. 2005). Approximately 42% of HCEs in vertebrates overlapped known exons, whereas >93% of HCEs in insect, worm, and yeast genomes were found to overlap exons. The most HCEs were found in 3′ UTRs especially those for regulatory genes (Siepel et al. 2005). Many of the HCEs associated with exons and 3′ UTRs seem to be enriched for RNA secondary structure and thus might represent unidentified coding regions (Siepel et al. 2005). Consistent with other studies, these authors noticed that stable gene deserts were well populated with HCEs and therefore this subclass of HCEs may represent noncoding elements that act as long-range regulators of developmental genes (Ovcharenko et al. 2005).

The fact that these sequences have been so highly conserved for tens or hundreds of millions of years is highly suggestive that they all are functional elements (Siepel et al. 2005). Indeed, Katzman et al. (2007) obtained evidence showing that human-rodent ultraconserved elements are under intense purifying selection. Findings from other studies also suggest that at least some of these HCEs act as long-range regulators of developmental genes or perform other gene regulatory functions (Nobrega et al. 2003; Woolfe et al. 2004; Bejerano et al. 2006; Pennacchio et al. 2006; Simons et al. 2006; Stephen et al. 2008). However, the function of most conserved elements remains unknown. These and other comparative genome studies have shown that eukaryotic genomes are well populated with these ultra- or HCEs, which, in turn, can be developed into a plethora of phylogenomic loci for tree of life studies.

The first sets of UCE- and AE-anchored loci were developed in animals, as the former was designed to obtain large multilocus datasets from

amniotes (Faircloth et al. 2012) while the latter was made for acquiring similar datasets from all vertebrates (Lemmon et al. 2012). In contrast to the original set of UCE loci, which were comprised of noncoding elements, the earliest set of AE loci primarily resided in coding regions of the genome (see Figure 2 in Lemmon et al. 2012; Lemmon and Lemmon 2013; Eytan et al. 2015). Despite any differences in the types of genomic sites found in UCE- and AE-anchored loci, both types of loci share the general and crucial properties of exhibiting extreme levels of DNA sequence conservation across wide phylogenetic distances (i.e., intraphyla) and offering potentially hundreds to thousands of genome-wide loci for phylogenomic studies.

If the sites within these elements are so highly conserved (i.e., >80% identical), then how can they be used to generate empirical gene trees? After all, sequences displaying little or no variation cannot be used to accurately infer gene trees. The answer is that the sites within the UCEs/AEs are themselves not used to reconstruct gene trees. Instead, the less-conserved flanking regions are used in phylogenomic analyses. Indeed, the level of variability increases with distance away from the core regions of UCE- and AE-anchored loci (see Figure 3 in Faircloth et al. 2012 and Figure 2 in Lemmon et al. 2012). The primary role of these HCEs is to serve as *anchors* to which oligonucleotide probes are hybridized during the data acquisition process (Chapter 7). Briefly, for each UCE- or AE-anchored locus, sequences are obtained from different individual genomes that include both the probe region (anchor location) as well as some amount of sequence from the flanking regions. Thus, the sites closest to the anchor location will be most conserved and thus can be used to reconstruct the deep nodes of a species tree while the more distant and variable sites are useful for inferring the relationships among recently diverged species (Faircloth et al. 2012; Lemmon et al. 2012). The anchors, therefore, represent the key elements that allow researchers to acquire orthologous sequences from lineages of organisms that diverged from each other hundreds of millions of years ago. Moreover, despite the fact that the anchor portion of these sequences may contribute little or no information to gene tree inferences, they are also useful for facilitating multiple sequence alignments among highly diverged species (Faircloth et al. 2012; Lemmon et al. 2012). What types of sites are in these flanking regions? Owing to the generally low or nonexistent levels of sequence conservation in UCE flanking regions, these sites are presumably nonfunctional and therefore more closely resemble anonymous loci. In contrast, the flanking regions of AE-anchored loci represent a mixture of coding, intron, and other sequences (see Figure 2 in Lemmon et al. 2012).

Many recent phylogenomic studies based on anchored loci used coalescent methods to infer species trees. One potential problem with this practice is that it is not yet clear how violating the neutrality assumption of these coalescent methods will affect the evolutionary inferences obtained from the datasets. Because all anchored loci include stretches of highly conserved noncoding or coding sites, which are very likely under purifying selection (Siepel et al. 2005; Katzman et al. 2007), the flanking sites will be subjected to indirect selection. As we saw earlier in this chapter, such selection is expected to diminish the genetic diversity of linked nonfunctional sites. This means that the gene trees representing the flanking regions of anchored loci will not only reflect the historical processes of mutation and genetic drift, but also that of direct or indirect natural selection. Accordingly, a set of anchored loci will yield a distribution of presumably independent gene trees that do not match coalescent expectations. McCormack et al. (2012) pointed out that selection on UCE-anchored loci may not adversely affect species tree inferences because selection would increase the rate of lineage sorting and thus the inferred gene trees would better match the topology of the true species tree. In support of this idea, Faircloth et al. (P. 721, 2012) noted that the distribution of inferred gene tree topologies based on a set of UCE-anchored loci for hominoids (human, chimpanzee, and gorilla) was similar to well-established expectations. Given this result, these authors concluded that UCE-anchored loci follow coalescent processes. Thus, despite anchored loci not being strictly neutral, they may still have the capability to generate the true species tree topology in coalescent-based analyses. This encouraging result should be verified by more studies especially those involving simulations.

The use of anchored loci for estimating historical demographic parameters (e.g., effective population size and historical gene flow), as practiced in some studies (e.g., Smith et al. 2013), may be more

problematic. This is because the loss of genetic diversity due to the effects of indirect selection is expected to lead to underestimates of effective population sizes, which are key parameters in phylogeography studies (McVicker et al. 2009; Costa et al. 2016). Anonymous loci on the other hand, are likely not impacted by direct or indirect selection and hence distributions of gene trees based on these loci should better reflect coalescent expectations. Until we know more about the effects of violating the neutrality assumption in multilocus coalescent analyses, it seems prudent to exercise some caution when interpreting phylogeographic results based on anchored UCE or AE loci (Costa et al. 2016).

REFERENCES

Amaral, F. R., S. V. Edwards, and C. Y. Miyaki. 2012. Eight anonymous nuclear loci for the squamate antbird (*Myrmeciza squamosa*), cross-amplifiable in other species of typical antbirds (Aves, Thamnophilidae). *Conserv Genet Resour* 4:645–647.

Arbogast, B. S., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu Rev Ecol Syst* 33:707–740.

Avise, J. C. 2000. *Phylogeography: The History and Formation of Species*. Cambridge: Harvard University Press.

Backström, N., W. Forstmeier, H. Schielzeth et al. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res* 20:485–495.

Baker, C. and S. Palumbi. 1994. Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* 265:1538–1539.

Baker, C. S., A. Perry, J. L. Bannister et al. 1993. Abundant mitochondrial DNA variation and worldwide population structure in humpback whales. *Proc Natl Acad Sci USA* 90:8239–8243.

Balakrishnan, C. N. and S. V. Edwards. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics* 181:645–660.

Balakrishnan, C. N., R. Ekblom, M. Völker et al. 2010. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biol* 8:29.

Ballard, J. W. O. and D. M. Rand. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. *Annu Rev Ecol Evol Syst* 36:621–642.

Begun, D. J. and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–520.

Bejerano, G., C. B. Lowe, N. Ahituv et al. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441:87–90.

Bejerano, G., M. Pheasant, I. Makunin et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Bensasson, D., D. X. Zhang, D. L. Hartl, and G. M. Hewitt. 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends Ecol Evol* 16:314–321.

Bertozzi, T., K. L. Sanders, M. J. Sistrom, and M. G. Gardner. 2012. Anonymous nuclear loci in non-model organisms: Making the most of high throughput genome surveys. *Bioinformatics* 28:1807–1810.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Brenner, S. 1998. Refuge of spandrels. *Curr Biol* 8:R669.

Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis.

Brown, W. M., M. George, and A. C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 76:1967–1971.

Brumfield, R. T., P. Beerli, D. A. Nickerson, and S. V. Edwards. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* 18:249–256.

Charlesworth, B. 2012. The effects of deleterious mutations on evolution at linked sites. *Genetics* 190:5–22.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141:1619–1632.

Chen, F. C. and W.-H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68:444–456.

Chenuil, A., T. B. Hoareau, E. Egea et al. 2010. An efficient method to find potentially universal population genetic markers, applied to metazoans. *BMC Evol Biol* 10:1.

Collura, R. V. and C. B. Stewart. 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature* 378:485–489.

Costa, I. R., F. Prosdocimi, and W. B. Jennings. 2016. In silico phylogenomics using complete genomes: A case study on the evolution of hominoids. *Genome Res* 26:1257–1267.

Crawford, N. G., B. C. Faircloth, J. E. McCormack, R. T. Brumfield, K. Winker, and T. C. Glenn. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett* 8:783–786.

Daguin, C., F. Bonhomme, and P. Borsa. 2001. The zone of sympatry and hybridization of *Mytilus edulis* and *M. galloprovincialis*, as described by intron length polymorphism at locus mac-1. *Heredity* 86:342–354.

Darst, C. R. and D. C. Cannatella. 2004. Novel relationships among hyloid frogs inferred from 12S and 16S mitochondrial DNA sequences. *Mol Phylogenet Evol* 31:462–475.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375.

Derti, A., F. P. Roth, G. M. Church, and C. Wu. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* 38:1216–1220.

Doolittle, W. F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 110:5294–5300.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.

Edwards, S. V. and P. Beerli. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54:1839–1854.

Edwards, S. V. and S. Bensch. 2009. Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Mol Ecol* 18:2930–2933.

Edwards, S. V., W. B. Jennings, and A. M. Shedlock. 2005. Phylogenetics of modern birds in the era of genomics. *Proc R Soc Lond B Biol Sci* 272:979–992.

Eisen, M. 2012. A neutral theory of molecular function. It is NOT junk, a blog about genomes, DNA, evolution, open science, baseball and other important things. http://www.michaeleisen.org/blog/?p=1172 (accessed October 16, 2015).

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Eytan, R. I., B. R. Evans, A. Dornburg et al. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. *BMC Evol Biol* 15:1.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol* 61:717–726.

Faircloth, B. C., L. Sorenson, F. Santini, and M. E. Alfaro. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS ONE* 8:e65923.

Fedorov, A., A. F. Merican, and W. Gilbert. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci USA* 99:16128–16133.

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics* 78:737–756.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland: Sinauer.

Felsenstein, J. 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol Biol Evol* 23:691–700.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst Biol* 19:99–113.

Friesen, V. L. 2000. Introns. In *Molecular Methods in Ecology*, ed. A. J. Baker, 274–294. Oxford: Blackwell.

Fu, Y.-X. and W.-H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Galtier, N., B. Nabholz, S. Glémin, and G. D. D. Hurst. 2009. Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Mol Ecol* 18:4541–4550.

Gillespie, J. 2004. *Population Genetics: A Concise Guide*, 2nd edition. Baltimore: The Johns Hopkins University Press.

Gottscho, A. D., S. B. Marks, and W. B. Jennings. 2014. Speciation, population structure, and demographic history of the Mojave Fringe-toed Lizard (*Uma scoparia*), a species of conservation concern. *Ecol Evol* 4:2546–2562.

Graur, D. and W.-H. Li. 2000. *Fundamentals of Molecular Evolution*, 2nd edition. Sunderland: Sinauer.

Graur, D., Y. Zheng, and R. B. Azevedo. 2015. An evolutionary classification of genomic function. *Genome Biol Evol* 7:642–645.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5:578–590.

Gray, G. S. and W. M. Fitch. 1983. Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol Biol Evol* 1:57–66.

Hare, M. P. 2001. Prospects for nuclear gene phylogeography. *Trends Ecol Evol* 16:700–706.

Hartl, D. L. and E. W. Jones. 2006. *Essential Genetics: A Genomics Perspective*, 4th edition. Sudbury: Jones and Bartlett Publishers.

Hassan, M., C. Lemaire, C. Fauvelot, and F. Bonhomme. 2002. Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. *Mol Ecol Notes* 2:334–340.

Hazkani-Covo, E., R. M. Zeller, and W. Martin. 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 6:e1000834.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. *Proc R Soc Lond B Biol Sci* 270:313–321.

Hess, C. M. and S. V. Edwards. 2002. The evolution of the major histocompatibility complex in birds. *Bioscience* 52:423–431.

Hillis, D. M. and S. K. Davis. 1986. Evolution of ribosomal DNA: Fifty million years of recorded history in the frog genus *Rana*. *Evolution* 40:1275–1288.

Hillis, D. M. and M. T. Dixon. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. *Q Rev Biol* 66:411–453.

Hudson, R. R. and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56:1557–1565.

Hudson, R. R. and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111:147–164.

Hudson, R. R. and N. L. Kaplan. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine. 2002. Inferring the root of a phylogenetic tree. *Syst Biol* 51:32–43.

Hurtley, S. 2012. No more junk DNA. *Science* 337:1581.

Innan, H. and W. Stephan. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* 165:2307–2312.

Jarvis, E. D., S. Mirarab, A. J. Aberer et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Jennings, W. B. 2016. On the independent loci assumption in phylogenomic studies. *bioRxiv* doi: http://dx.doi.org/10.1101/066332.

Jennings, W. B. and S. V. Edwards. 2005. Speciational history of Australian Grass Finches (*Poephila*) inferred from thirty gene trees. *Evolution* 59:2033–2047.

Jennings, W. B., E. R. Pianka, and S. Donnellan. 2003. Systematics of the lizard family Pygopodidae with implications for the diversification of Australian temperate biotas. *Syst Biol* 52:757–780.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. *Genetics* 123:887–899.

Karl, S. A. and J. C. Avise. 1993. PCR-based assays of mendelian polymorphisms from anonymous single-copy nuclear DNA: Techniques and applications for population genetics. *Mol Biol Evol* 10:342–361.

Katzman, S., A. D. Kern, G. Bejerano et al. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915–915.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.

King, J. L. and T. H. Jukes. 1969. Non-Darwinian evolution. *Science* 164:788–798.

Kocher, T. D., W. K. Thomas, A. Meyer et al. 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc Natl Acad Sci USA* 86:6196–6200.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* 31:241–247.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304:412–417.

Kuhner, M. K., P. Beerli, J. Yamato, and J. Felsenstein. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156:439–447.

Lanier, H. C. and L. L. Knowles. 2012. Is recombination a problem for species-tree analyses? *Syst Biol* 61:691–701.

Lee, J. Y. and S. V. Edwards. 2008. Divergence across Australia's Carpentarian barrier: Statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution* 62:3117–3134.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol* 61:727–744.

Lemmon, A. R. and E. M. Lemmon. 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Syst Biol* 61:745–761.

Lemmon, E. M. and A. R. Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst* 44:99–121.

Lessa, E. P. 1992. Rapid surveying of DNA sequence variation in natural populations. *Mol Biol Evol* 9:323–330.

Lewontin, R. C. and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54:595–609.

Li, C., G. Ortí, G. Zhang, and G. Lu. 2007. A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol Biol* 7:44.

Li, C., J.-J. M. Riethoven, and L. Ma. 2010. Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evol Biol* 10:90.

Lindblad-Toh, K., M. Garber, O. Zuk et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482.

Martin, D. P., P. Lemey, M. Lott, V. Moulton, D. Posada, and P. Lefeuvre. 2010. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.

Maynard Smith, J. and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* 23:23–35.

McCormack, J. E., B. C. Faircloth, N. G. Crawford, P. A. Gowaty, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res* 22:746–754.

McCormack, J. E., M. G. Harvey, B. C. Faircloth, N. G. Crawford, T. C. Glenn, and R. T. Brumfield. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS ONE* 8:e54848.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

McVicker, G., D. Gordon, C. Davis, and P. Green. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5: e1000471.

Meader, S., C. P. Ponting, and G. Lunter. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* 20:1335–1343.

Mindell, D. P. and R. L. Honeycutt. 1990. Ribosomal RNA in vertebrates: Evolution and phylogenetic applications. *Annu Rev Ecol Syst* 21:541–566.

Mindell, D. P. and A. Meyer. 2001. Homology evolving. *Trends Ecol Evol* 16:434–440.

Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees. *Evolution* 49:718–726.

Moritz, C., T. E. Dowling, and W. M. Brown. 1987. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. *Annu Rev Ecol Syst* 18:269–292.

Moritz, C. and D. M. Hillis. 1996. Chapter 1. Molecular systematics: Context and controversies. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 1–13. Sunderland: Sinauer.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–942.

Nielsen, R., M. J. Hubisz, and A. G. Clark. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–2382.

Nobrega, M. A., I. Ovcharenko, V. Afzal, and E. M. Rubin. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302:413–413.

Ohno, S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol* 23:366–370.

Ohresser, M., P. Borsa, and C. Delsert. 1997. Intron-length polymorphism at the actin gene locus mac-1: A genetic marker for population studies in the marine mussels *Mytilus galloprovincialis* Lmk. and *M. edulis* L. *Mol Mar Biol Biotechnol* 6:123–130.

Ohta, T. and M. Kimura. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68:571–580.

O'Neill, E. M., R. Schwartz, C. T. Bullock et al. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Mol Ecol* 22:111–129.

Ovcharenko, I., G. G. Loots, M. A. Nobrega, R. C. Hardison, W. Miller, and L. Stubbs. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res* 15:137–145.

Palazzo, A. F. and T. R. Gregory. 2014. The case for junk DNA. *PLoS Genet* 10:e1004351.

Palazzo, A. F. and E. S. Lee. 2015. Non-coding RNA: What is functional and what is junk? *Front Genet* 6:1–11.

Palumbi, S. R. 1996. Chapter 7. Nucleic acids II: The polymerase chain reaction. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 205–247. Sunderland: Sinauer.

Palumbi, S. R. and C. S. Baker. 1994. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Mol Biol Evol* 11:426–435.

Pennacchio, L. A., N. Ahituv, A. M. Moses et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502.

Pennisi, E. 2012. ENCODE project writes eulogy for junk DNA. *Science* 337:1159–1161.

Philippe, H. and M. Blanchette. 2007. Overview of the first phylogenomics conference. *BMC Evol Biol* 7(Suppl 1)S1.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst* 36:541–562.

Pluzhnikov, A. and P. Donnelly. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–1262.

Ponting, C. P. and R. C. Hardison. 2011. What fraction of the human genome is functional? *Genome Res* 21:1769–1776.

Portik, D. M., P. L. Wood Jr, J. L. Grismer, E. L. Stanley, and T. R. Jackman. 2011. Identification of 104 rapidly-evolving nuclear protein-coding markers for amplification across scaled reptiles using genomic resources. *Conserv Genet Resour* 4:1–10.

Prum, R. O., J. S. Berv, A. Dornburg et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569.

Prychitko, T. M. and W. S. Moore. 1997. The utility of DNA sequences of an intron from the β-fibrinogen gene in phylogenetic analysis of woodpeckers (Aves: Picidae). *Mol Phylogenet Evol* 8:193–204.

Randi, E. 2000. Mitochondrial DNA. In *Molecular Methods in Ecology*, ed. A. J. Baker, 136–167. Oxford: Blackwell.

Rannala, B. and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.

Reich, D. E., M. Cargill, S. Bolk et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204.

Reilly, S. B., S. B. Marks, and W. B. Jennings. 2012. Defining evolutionary boundaries across parapatric ecomorphs of Black Salamanders (*Aneides flavipunctatus*) with conservation implications. *Mol Ecol* 21:5745–5761.

Robertson, A. and W. G. Hill. 1983. Population and quantitative genetics of many linked loci in finite populations. *Proc R Soc Lond B Biol Sci* 219:253–264.

Rosenblum, E. B., N. M. Belfiore, and C. Moritz. 2007a. Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. *Mol Ecol Notes* 7:113–116.

Rosenblum, E. B., M. J. Hickerson, and C. Moritz. 2007b. A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution* 61:2971–2985.

Rosenblum, E. B. and J. Novembre. 2007. Ascertainment bias in spatially structured populations: A case study in the eastern fence lizard. *J Hered* 98:331–336.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Saint, K. M., C. C. Austin, S. C. Donnellan, and M. N. Hutchinson. 1998. C-mos, a nuclear marker useful for squamate phylogenetic analysis. *Mol Phylogenet Evol* 10:259–263.

Santiago, E. and A. Caballero. 1998. Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics* 149:2105–2117.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto. 2009. Pervasive natural selection in the *Drosophila* genome. *PLoS Genet* 5:e1000495.

Sequeira, F., N. Ferrand, and D. J. Harris. 2006. Assessing the phylogenetic signal of the nuclear β-Fibrinogen intron 7 in salamandrids (Amphibia: Salamandridae). *Amphibia-Reptilia* 27:409–418.

Shaffer, H. B. and R. C. Thomson. 2007. Delimiting species in recent radiations. *Syst Biol* 56:896–906.

Siepel, A., G. Bejerano, J. S. Pedersen et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.

Simons, C., M. Pheasant, I. V. Makunin, and J. S. Mattick. 2006. Transposon-free regions in mammalian genomes. *Genome Res* 16:164–172.

Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2013. Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. *Syst Biol* doi: 10.1093/sysbio/syt061.

Smith, M. F., W. K. Thomas, and J. L. Patton. 1992. Mitochondrial DNA-like sequence in the nuclear genome of an akodontine rodent. *Mol Biol Evol* 9:204–215.

Sorenson, M. D. and T. W. Quinn. 1998. Numts: A challenge for avian systematics and population biology. *Auk* 115:214–221.

Stapley, J., T. R. Birkhead, T. Burke, and J. Slate. 2008. A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics* 179:651–667.

Stephan, W. 2010. Genetic hitchhiking versus background selection: The controversy and its implications. *Philos Trans R Soc Lond B Biol Sci* 365:1245–1253.

Stephan, W., T. H. Wiehe, and M. W. Lenz. 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theor Popul Biol* 41:237–254.

Stephen, S., M. Pheasant, I. V. Makunin, and J. S. Mattick. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* 25:402–408.

Swofford, D. L. and G. J. Olsen. 1990. Phylogeny reconstruction. In *Molecular Systematics*, eds. D. M. Hillis and C. Moritz, 411–501. Sunderland: Sinauer.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Chapter 11. Phylogenetic inference. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 407–514. Sunderland: Sinauer.

Takahata, N. 1993. Allelic genealogy and human evolution. *Mol Biol Evol* 10:2–22.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.

Tenesa, A., P. Navarro, B. J. Hayes et al. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17:520–526.

Thomson, G. 1977. The effect of a selected locus on linked neutral loci. *Genetics* 85:753–788.

Thomson, R. C., I. J. Wang, and J. R. Johnson. 2010. Genome-enabled development of DNA markers for ecology, evolution and conservation. *Mol Ecol* 19:2184–2195.

Thorne, J. L. and H. Kishino. 2005. Chapter 8. Estimation of divergence times from molecular sequence data. In *Statistical Methods in Molecular Evolution*, ed. R. Nielsen, 233–256. New York: Springer.

Vences, M., M. Thomas, A. Van der Meijden, Y. Chiari, and D. R. Vieites. 2005. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Front Zool 2:5.

Vigilant, L., R. Pennington, H. Harpending, T. D. Kocher, and A. C. Wilson. 1989. Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci USA* 86:9350–9354.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507.

Wakeley, J. 2009. *Coalescent Theory: An Introduction* (Vol. 1). Greenwood Village: Roberts & Company Publishers.

Wakeley, J. and J. Hey. 1997. Estimating ancestral population parameters. *Genetics* 145:847–855.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. 2001. The discovery of single-nucleotide polymorphisms—And inferences about human demographic history. *Am J Hum Genet* 69:1332–1347.

Warren, W. C., D. F. Clayton, H. Ellegren et al. 2010. The genome of a songbird. *Nature* 464:757–762.

Watson, J. D., T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick. 2014. *Molecular Biology of the Gene*, 7th edition. New York: Pearson Education, Inc.

Weibel, A. C. and W. S. Moore. 2002. A test of a mitochondrial gene-based phylogeny of woodpeckers (genus *Picoides*) using an independent nuclear gene, β-fibrinogen intron 7. *Mol Phylogenet Evol* 22:247–257.

Wiehe, T. H. and W. Stephan. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol Biol Evol* 10:842–854.

Wilson, A. C., R. L. Cann, S. M. Carr et al. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol J Linn Soc Lond* 26:375–400.

Woerner, A. E., M. P. Cox, and M. F. Hammer. 2007. Recombination-filtered genomic datasets by information maximization. *Bioinformatics* 23:1851–1853.

Woese, C. R. and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–5090.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci USA* 87:4576–4579.

Woolfe, A., M. Goodson, D. K. Goode et al. 2004. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3:e7.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–1823.

Yang, Z. 2006. *Computational Molecular Evolution* (Vol. 21). Oxford: Oxford University of Press.

Zhang, D. X. and G. M. Hewitt. 1996. Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends Ecol Evol* 11:247–251.

Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci USA* 77:2158–2162.

Zink, R. M. and G. F. Barrowclough. 2008. Mitochondrial DNA under siege in avian phylogeography. *Mol Ecol* 17:2107–2121.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, eds. V. Bryson and H. J. Vogel, 97–166. New York: Academic Press.