æ

4

THE NONADAPTIVE FORCES OF EVOLUTION

20 May 2014

Although natural selection plays a major role in the evolution of many traits, three additional factors determine the patterns of genetic variation within and among populations. We refer to these factors – mutation, recombination, and random genetic drift – as the **nonadaptive forces of evolution** because their operation is generally independent of the specific selective factors operating on the extrinsic phenotypes of individuals. Migration (briefly touched upon in Chapters 2 and 3) might be added to this list, although we regard this added complexity as being independent of the internal genetic machinery of a population. As will become clear in the following chapters, the three nonadaptive forces together comprise the population-genetic environment, which defines the paths of evolutionary change that are open vs. closed to natural selection.

Knowledge of the magnitude of the nonadaptive forces of evolution should be sufficient to arrive at a full description of the dynamics of allele/gamete-frequency change within populations in the absence of external forces of selection. Moreover, this logic works in reverse – under certain assumptions, observed patterns of variation in neutral genomic regions can be used to infer the magnitude of the evolutionary forces responsible for such patterning.

The goals of this chapter are, therefore, three-fold. First, we will consider how observations on putatively neutral molecular markers can be used to estimate rates of mutation, recombination, and random genetic drift. Second, we will summarize the existing data resulting from such analyses, providing information that will play a central role in applications presented in subsequent chapters. As a consequence of the recent emergence of new technologies for high-throughput genomic sequencing, this is a rapidly developing area that will undoubtedly experience additional refinements in the near future. Finally, having shown that the intensity of the nonadaptive forces of evolution vary by orders of magnitude among species in fairly predictable

manners, we will summarize existing theory that helps explain such patterns of variation.

Although our ultimate desire is to obtain accurate estimates of the individual forces of mutation, recombination, and drift, as will be seen below, it is often much easier to obtain ratios of these features than to measure them individually. Fortunately, this is not always an undesirable situation, for as we have seen in Chapter 2, in the absence of selection, the ratio of the power of mutation (u) and the power of drift ($1/2N_e$) defines the level of heterozygosity in a population, and the ratio of the recombination rate (c) and the power of drift defines the magnitude of linkage disequilibrium. Therefore, before summarizing the approaches for estimating N_e, u, and c separately, we will first consider methods for estimating the composite population parameters $\theta = 4N_e u$ and $\rho = 4N_e c$. As will be seen below, accurate estimates of N_e are particularly difficult to achieve directly, especially for large populations. However, by using combined estimates of θ, ρ, u, and/or c, approximate measures of long-term N_e are sometimes possible.

Throughout this chapter, we will assume that we are dealing with molecular markers known in advance to be behaving in an effectively neutral fashion. Numerous methods to test this hypothesis will be discussed in Chapters 8 and 9. We will largely focus on measures at the level of individual nucleotide sites, as it is now routine to obtain large quantities of DNA-sequence data, and per-nucleotide site measures are readily extrapolated to larger units of analysis such as genetic loci. Thus, u and c will respectively denote mutation and recombination rates per nucleotide site.

RELATIVE POWER OF MUTATION AND GENETIC DRIFT

In Chapter 2, it was demonstrated that if the forces of drift and mutation remain constant for a sufficiently long time, the level of heterozygosity at a neutral nucleotide site (with four possible allelic states) will stochastically wander around an expected equilibrium value of $\sim 12N_e u/(3 + 16N_e u)$, where u is the mutation rate per gamete per nucleotide site (assuming all nucleotides mutate at the same rate). As will be seen below, the average heterozygosity per neutral nucleotide site is far below 1.0 in all phylogenetic groups, so the preceding expression is closely approximated by $4N_e u$. This particular relationship has great practical utility. Because $2u$ is the mutation rate per site per diploid genome, $4N_e u$ is equivalent to the ratio of the power of mutation per diploid individual to the power of random genetic drift, $1/(2N_e)$. (For haploid species, the expected nucleotide diversity at a neutral site is $2N_e u$).

Nucleotide Diversity

Suppose a population sample of n random sequences has been obtained for a particular genomic region. In principle, such a stretch of DNA might consist of intronic or intergenic sequence or of the subset of **silent** (**synonymous**) sites in one or more coding regions. Letting k_{ij} be the number of site-specific differences between

observed sequences i and j, and L be the number of sites per sequence, the average fraction of pairwise differences between the sampled sequences,

$$\widehat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{\substack{i=1 \\ j>i}}^{n} k_{ij}/L \tag{4.1}$$

yields a heterozygosity-based estimate of $\theta = 4N_e u$ (Tajima 1983). (This estimate is frequently called the **Tajima estimator**, and is often denoted by π in the literature). However, if we are to confidently use $\widehat{\theta}_\pi$ as an estimator of θ, aside from knowing whether the assumptions of neutrality and equilibrium are valid, it is critical to know the sampling variance of $\widehat{\theta}_\pi$. Such variance results from two sources of uncertainty.

First, heterozygosity is subject to **evolutionary variance**, which results from the natural fluctuations of nucleotide frequencies over time generated by the stochastic forces of mutation and drift (Chapter 2). Although this source of variation is not easily observed directly, assuming a population in drift-mutation equilibrium, the expected evolutionary variance of the true population value of θ based on L independent (effectively unlinked) sites is $\simeq (\theta/3)[(2\theta/3) + (1/L)]$ (Tajima 1983). This source of variance is intrinsic to the features of the population, independent of the sample taken. Second, **sampling variance** results from the reliance of estimates of θ_π on a finite number of sampled sequences. For an equilibrium population, this variance is $\simeq \{2\theta/[3(n-1)]\}\{[(2n+3)\theta/3n] + (1/L)\}$, where n is the number of sequences sampled per site (Tajima 1983).

Summing over these two sources of variance, for sites in stochastic drift-mutation equilibrium, the expected total variance of heterozygosity-based estimates of θ is

$$\sigma^2(\widehat{\theta}_\pi) \simeq \frac{\theta}{3(n-1)} \left(\frac{2(n^2+n+3)\theta}{3n} + \frac{n+1}{L} \right) \tag{4.2}$$

(Pluzhnikov and Donnelly 1996). With increasing numbers of sampled alleles per site, i.e., as $n \to \infty$, the total variance of estimates of θ based on nucleotide diversity approaches a minimum equal to the evolutionary variance. Even with an enormous amount of sequence per individual (large L), the sampling coefficient of variation of $\widehat{\theta}_\pi$ is $\simeq \sqrt{2/9} \simeq 0.47$. Adding more individuals or sites to a survey will not alter this minimum.

It is worth reemphasizing that Equation 4.2 is an appropriate estimator of the variance of a nucleotide-diversity estimate only if the latter is based on neutral sites in drift-mutation equilibrium (the standard neutral model). Even for assuredly neutral sites, this expression will not apply for *nonequilibrium situations,* e.g., populations that have experienced relatively recent expansions or contractions. Under the latter conditions, the variance of heterozygosity must be evaluated more directly from the spectrum of allele frequencies across all sites and higher-order moments of them. A number of related technical issues are covered and general expressions derived in Nei and Roychoudhury (1974), Nei (1978), Nei and Tajima (1981a, 1983), Nei and Jin (1989), and Lynch and Crease (1990).

Finally, with high-throughput sequencing now being routine for entire diploid genomes, it is possible to estimate the average nucleotide diversity over millions of putatively neutral sites, yielding per-individual measures with near zero sampling variance. Because most pairs of sites are on different chromosomes, a full

survey of even a single individual from a random-mating population should provide a very accurate description of the average per-site diversity across the entire population. Moreover, when a survey of two random individuals is possible, the covariance of heterozygosity within sites provides a direct estimate of the evolutionary variance of heterozygosity among sites, which as noted above should closely approximate $(\theta/3)[(2\theta/3) + (1/L)]$ under the assumption of drift-mutation equilibrium (Lynch 2008a). These observations are now quite salient, as Pluzhnikov and Donnelly (1996) have shown that for a fixed amount of resources for sequencing Ln total bases, the optimal strategy for obtaining minimal-variance estimates of θ is generally to sample no more than two or so individuals, putting the effort instead into sampling more sites, i.e., maximizing L at the expense of n.

Number of Segregating Sites

Although nucleotide diversity is the most transparent means of estimating θ, it is by no means the only or even the most efficient approach. Watterson (1975) pointed out an alternative statistical measure of allelic diversity – the total number of segregating sites (S) in the region analyzed over the full set of n sequences. Because a segregating site is any nucleotide position that harbors two or more variants, S clearly increases with the length L of the sequence and the number of individuals assayed, but Watterson (1975) showed that under the assumptions of neutrality and drift-mutation equilibrium, an unbiased estimator of the per-site parameter $\theta = 4N_e u$ is

$$\widehat{\theta}_S = S/(La_n) \tag{4.3a}$$

where

$$a_n = \sum_{j=1}^{n-1} 1/j \tag{4.3b}$$

A central point here is that when the nucleotide sites surveyed are neutral and in drift-mutation equilibrium, like the Tajima estimator (Equation 4.1), the **Watterson equation** provides a separate estimate of θ. In Chapters 8 and 9, we will see that when the assumptions of neutrality and/or equilibrium are violated, the values of $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ deviate from each other in ways that yield insight into past population-genetic processes.

The sampling variance for the Watterson estimator, analogous to Equation 4.2 and again under the assumptions of neutrality and equilibrium, is

$$\sigma^2(\widehat{\theta}_S) \simeq \frac{\theta}{a_n}\left(\frac{\theta b_n}{a_n} + \frac{1}{L}\right) \tag{4.4a}$$

where

$$b_n = \sum_{j=1}^{n-1} 1/j^2 \tag{4.4b}$$

For sample sizes smaller than ten, the Tajima and Watterson estimators have similar expected sample standard deviations, but with larger n, the latter can be up to two-fold smaller than the former, although there is little to be gained with either

approach once n exceeds 50 or so (Figure 4.1). It should, however, be emphasized that both Equations 4.2 and 4.4a were derived under the assumptions of sequences experiencing *negligible recombination*. The necessary modifications to allow for intragenic recombination, derived in Pluzhnikov and Donnelly (1996; their Equations 6 and 7), play a role in some methods for estimating the population recombination rate, as described in the following section.

-Insert Figure 4.1 Here-

One significant issue that arises with the use of S to estimate θ in the modern era of high-throughput sequencing involves the introduction of upward bias from sequencing errors. With large numbers of sites and individuals, errors will inevitably appear as singletons but nonetheless enter the estimate of S. Such effects can be quite deceptive in population-genetic analyses because rare alleles are expected to be common under the neutral hypothesis. Johnson and Slatkin (2008), Kang and Marjoram (2011), and Keightley and Halligan (2011) have suggested methods for eliminating the bias from S when an accurate estimate of the sequencing-error rate is available. An alternative approach relaxes this constraint by estimating the error rate from the data themselves (Lynch 2009).

Alternative Approaches

Felsenstein (1992) pointed out that neither of the above approaches are likely to provide the most efficient estimates of θ (i.e., to yield estimates with minimum sampling variance), as they do not utilize all of the information in the sample of sequences. In particular, both approaches ignore the genealogical relationships of sequences (i.e., the coalescent structure of the sample), although as shown in Chapter 2, under neutrality the expected contribution to variation from each genealogical branch can be expressed in terms of θ.

To evaluate how much improvement might be achieved by exploiting such information, Fu and Li (1993a) derived a maximum-likelihood estimator of θ for the extreme situation in which one knows with certainty the genealogical relationships of the sequences and the numbers of mutations and generations on each branch of the genealogy. The expected sampling variance of this estimator is

$$\sigma^2(\widehat{\theta}_{ML}) = \frac{\theta}{a_n}\left(\frac{\theta a_n}{n-1} + \frac{1}{L}\right) \tag{4.5}$$

Comparison of this expression and Equations 4.2 and 4.4 illustrates that there is substantial room for improvement in the estimation of θ over the traditional heterozygosity and segregating-sites methods, provided the number of sequences exceeds five or so, and assuming a reasonably accurate gene genealogy can be obtained (Figure 4.1).

Gene genealogies cannot be constructed without error. However, using information on the expected coalescence times of samples of neutral sequences, Fu (1994a,b) developed several generalized least-squares estimators that account for the sampling

variances and covariances of mutations on different branch segments. Several of these estimators, which utilize the concepts of the site-frequency spectrum (Fu 1995; Li and Fu 1999; see Chapter 2), asymptotically perform in a near optimal manner as the sample size increases, again provided the sites are neutral and the population is in drift-mutation equilibrium.

As one or both of the latter two assumptions (neutrality and equilibrium) are likely to be violated to unknown degrees in many natural settings, an estimator with minimum sensitivity to both problems would be highly useful. In fact, just such an approach can be extrapolated from Watterson's estimator (4.3). The basis for this strategy follows from the property that for neutral alleles, in an equilibrium population, the number of derived single-nucleotide variants found j times in a sample of size n, S_j, has expected value $L\theta/j$ (Watterson 1975; Fu 1995). Because the total number of segregating sites, S, has expected value $L\theta a_n$, it follows that Watterson's estimator is equivalent to an average of estimates of θ, with each contributor being based on a class of variants weighted by the inverse of the number of observations.

The simplest estimate of θ, based only on singletons ($j = 1$), is then

$$\widehat{\theta}_1 = S_1/L \tag{4.6a}$$

which is also equivalent to the number of mutations (per site) on the external branches of a gene genealogy (Fu and Li 1993b). Such an estimator is attractive for two reasons. First, the singletons in a sample are a function of the very recent past, especially when the overall sample size is large, and hence are not expected to be influenced by distant periods of population-size change. Second, because the dynamics of rare alleles are primarily governed by the drift process, singleton frequencies are expected to most closely reflect the pattern expected under neutrality even when such mutations are non-neutral (Messer 2009). The sampling variance of the singleton-based estimator is

$$\sigma^2(\widehat{\theta}_1) = \frac{\theta}{n}\left(\frac{n-1}{L} + \frac{\theta[2a_n(n-1)-1]}{n}\right) \tag{4.6b}$$

Considering just the sampling variance of the estimators of θ to this point, as $n \to \infty$, those for $\widehat{\theta}_W$ and $\widehat{\theta}_{ML}$ are $\theta/(15.4L)$, whereas that for $\widehat{\theta}_\pi$ is $\theta/(3L)$, and that for $\widehat{\theta}_1$ is θ/L. Thus, although the singleton-based estimator is likely to have the smallest amount of bias, a focus on only a fraction of the segregating sites results in higher sampling variance.

Empirical Observations

Estimates of θ, mostly derived as silent-site heterozygosity from protein-coding genes using Equation 4.1, have been summarized for a wide range of species by Lynch (2007)and Leffler et al. (2012). Across a diverse assemblage of > 100 eukaryotic and prokaryotic species, there is an inverse relationship between organism size and θ_π, with estimates for prokaryotes falling in the broad range of 0.007 to 0.388, with an average value of 0.104 (and a large standard deviation of 0.111). The average values for unicellular eukaryotes (mean = 0.057, SD = 0.078) and invertebrates

(mean = 0.026, SD = 0.015) are 50 to 75% lower, and estimates for land plants (mean = 0.015, SD = 0.013) and vertebrates (mean = 0.004, SD = 0.003) are still smaller. Because the numbers of independent studies contributing to these estimates are in the range of 15 to 50, the cited means should be quite reliable (with some caveats given below), but because of sampling error at the gene, individual, and population levels, the standard deviations must be upwardly biased with respect to that expected from true evolutionary variance.

For both of the unicellular groups, silent-site heterozygosity measures are likely to be downwardly biased estimators of $4N_eu$ ($2N_eu$ for haploids) for at least two reasons. First, most recorded studies of microbial species are derived from surveys of pathogens, whose N_e may be abnormally low because of the restricted distributions of their multicellular host species. Second, silent-site variation will underestimate the neutral expectation if such sites experience some form of purifying selection. Such conditions can arise for a variety of reasons: 1) translation-associated selection when certain tRNAs have higher affinities for certain alternative codons (**codon bias** resulting from differential tRNA abundance and/or physical features); 2) selection on sites involved in splice-junction identification for species with introns; 3) secondary selection against codons that are one mutational step from termination codons; and 4) inhibition of double-strand break repair between highly divergent alleles. The molecular biological underpinnings of some of these factors, as well as their potential population-genetic consequences, are reviewed in Lynch (2007). Because such selection is expected to be quite weak, it will be most effective in populations with very large N_e, and the central conclusion is that although θ_π may underestimate $4N_eu$ ($2N_eu$ for haploids) in some microbial species by as much as tenfold, the bias may be minor in multicellular eukaryotes. However, many uncertainties remain, and we return to the topic in Chapter 8.

With these caveats in mind, the existing data make a compelling statement with respect to the relative power of mutation and random genetic drift – in essentially no species is there evidence that the former exceeds the latter (as this would cause $4N_eu > 1$), and in large multicellular land plants and vertebrates, the ratio is almost always on the order of 0.03 or much smaller. Thus, drift appears to be a more powerful force than mutation at the nucleotide level in all species, except perhaps the smallest microbes. As the absolute population sizes of many species (certainly microbes) can exceed $1/u$ by orders of magnitude (see below), these observations clearly support the idea introduced in Chapter 3 that N_e is usually substantially smaller than the actual number of reproductive individuals in a population, and that in especially large populations, this is largely a consequence of selection on linked sites. As introduced in Chapter 3 and detailed in Chapter 8, much of this reduction may arise from selection on linked sites.

RELATIVE POWER OF RECOMBINATION AND GENETIC DRIFT

As will be seen in subsequent chapters, recombination plays an important role in evolution because the physical scrambling of linked genes increases the ability of natural selection to promote or eliminate mutations on the basis of their individ-

ual effects. On the other hand, high rates of recombination can often inhibit the establishment of pairs of mutations with favorable epistatic effects.

Two general approaches provide insight into the level of recombination per physical distance along chromosomes. Genetic maps, generally derived from controlled crosses, are based on observations on the frequency of meiotic crossovers between informative markers (LW Chapter 14), whereas studies of linkage disequilibrium (LD) in natural populations use the theoretical concepts introduced in Chapter 2 to indirectly infer the relative magnitudes of the joint forces of random genetic drift and recombination. High-density genetic maps have the power to yield accurate estimates of average recombination rates over fairly long physical distances (usually with markers being separated by millions of nucleotide sites, which typically corresponds to $> 1\%$ recombination per generation), but because they are typically outcomes of many thousands of generations, patterns of LD have the potential to reveal much more refined views (kilobase scale) of the recombinational landscape. For a mapping cross involving n gametes with a recombination frequency r between sites, the expected number of recombinants is nr, so for sufficiently close sites, the typical outcome will be a complete absence of recombinants. On the other hand, if n random chromosomes are sampled from a natural population with a mean coalescence time of $\bar{t} = 2N_e$ generations (Chapter 2), the expected number of recombination events is $2\bar{t}nr = 4N_enr$.

Recall from Chapter 2 that $\rho_L = 4N_er$ is the effective number of recombination events between sites per generation at the entire population level, which is also equivalent to the ratio of the power of recombination to the power of drift. Just as the amount of segregating variation at neutral sites provides insight into the population mutation rate $\theta = 4N_eu$, the amount of standing LD is a function of the population recombination rate. Although a wide variety of methods for estimating ρ have been proposed, the challenges to obtaining accurate measures are substantial. The markers employed must not only have at least moderate frequencies (to ensure accurate estimates of gamete frequencies and reasonable likelihoods of observing recombination events), but behave neutrally (to ensure the validity of the application of drift-recombination theory). Moreover, most of the proposed estimators rely on the assumption of drift-mutation-recombination equilibrium, while also suffering from very high sampling variance.

Number of Recombinational Events in a Sample of Alleles

We start with a description of methods involving shorts spans of DNA, e.g., single genes or pairs of adjacent genes with **phased haplotypes** (i.e., with complete sequences available for each of the two haplotypes within diploid individuals). Chromosomal regions of such small size will often have $c \ll 0.01$, and hence no chance of revealing recombinants in most mapping crosses. While population-level analyses can aid in the detection of historical recombination events, the power here is also limited. Crossover events only leave a trace if they involve pairs of doubly-heterozygous chromosomes, and there is no way to directly determine whether multiple recombinants in a sample are a result of parallel recombination events or intact descendants of the same events. Thus, to obtain unbiased estimates of ρ, we require a method

for converting the *observed* number of recombinant events in a sample to the *actual* number that must have occurred (R).

For neutral sites separated by L nucleotides, assuming a population in drift-mutation-recombination equilibrium, the expected value of R in a sample of n sequences is equal to $\rho L a_n$, where $\rho = 4N_c$ is the recombination parameter for adjacent sites (where ρ is again on a distance scale of single nucleotide sites), and a_n is given by Equation 4.3b (Hudson and Kaplan 1985). Rearranging, a potential estimator for ρ is

$$\widehat{\rho} = \widehat{R}/(La_n) \tag{4.7}$$

where $\widehat{R}$ is the estimated number of recombinational events that have occurred between the two sites in the history of the sample. Note the similarity of the form of this expression to that relating the number of segregating mutations to θ (Equation 4.3a).

The primary impediment to applying this expression is the estimation of R. One approach, proposed by Hudson and Kaplan (1985), starts with the **four-gamete test**, which asserts that any pair of heterozygous sites exhibiting four gametic haplotypes must reflect the prior action of at least one recombination event, assuming an absence of parallel mutations at the same sites in the history of the sample. Under this model, starting with a fixed gamete of the form **AB**, a single mutation will create either an **aB** or **Ab** gamete, resulting in two gametic types in the population, which are noninformative because a novel gamete cannot result from recombination with the ancestral haplotype. One of these gametic types might eventually go to fixation, recreating the initial scenario of double homozygosity. However, if prior to fixation a mutation arises at the remaining homozygous site, there will be three haplotypes (**AB**, **Ab**, and **aB**), with the fourth type (**ab**) arising only by subsequent recombination (in the absence of recurrent mutation). Judiciously applying this criterion to all pairs of segregating sites in a sample of sequences and ensuring that the same event is not counted more than once, it is possible to estimate $R_{\min}$, the minimum number of crossover events in the history of the sample (Hudson and Kaplan 1985). More complex approaches attempt to derive information from the complete haplotype structure in a sample (Myers and Griffiths 2003; Liu and Fu 2008).

In principle, with knowledge of the expected fraction of detectable recombination events, one could extrapolate the observed $R_{\min}$ to an estimate of the actual value R. Assuming conditions of drift-mutation equilibrium, Stephens (1986) found the lower and upper bounds to the fraction of random recombination events giving rise to observable, nonparental haplotypes,

$$d_{r,\min} = 1 - [2\ln(1+\Theta)]/\Theta + [1/(1+\Theta)] \tag{4.8a}$$

$$d_{r,\max} = 1 - [2(1-e^{-\Theta})]/\Theta + e^{-\Theta} \tag{4.8b}$$

where $\Theta = \theta L$ is the population mutation rate for the stretch of DNA being surveyed, with L being the maximum distance between segregating sites in the sample. These two limits are respectively approached as $c \to 0.0$ (complete linkage) and 0.5 (free recombination). As θ is generally on the order of 0.001 to 0.01 for neutral sites, unless the segments being analyzed have lengths in excess of 1000 nucleotides, the

majority of recombination events will simply reproduce parental gamete types, and hence not be scored as recombinants (Figure 4.2).

-Insert Figure 4.2 Here-

Given an estimate of Θ, Equations 4.8a,b can be used to approximate the total number of recombination events in the sample as $R_{\min}/\,\overline{d}_r$, where $\overline{d}_r$ is the average of $d_{r,\min}$ and $d_{r,\max}$. However, even this approach is not fully adequate because only a subset of the recombinant gametes that are nonparental with respect to markers are also novel with respect to the entire population, i.e., the fraction of uniquely detectable recombination events is even lower than suggested by Equations 4.7a,b.

An empirical approach to this problem was suggested by Zietkiewicz et al. (2003; see also Lefebre and Labuda 2008). Letting p_i denote the frequency of the ith haplotype in a sample, an estimator for the fraction of detectable (but not necessarily unique) recombinant alleles is

$$\widehat{d}_r = \sum_{\substack{i=1 \\ j>i}}^{L} 2p_i p_j L_{\max,ij}/L \tag{4.9}$$

where $L_{\max,ij}$ is the distance between the maximally separated heterozygous sites in the ijth comparison. Through simulations, one can establish the fraction of potentially informative recombination events that would indeed produce novel haplotypes in the sample, thereby reducing $\widehat{d}_r$ to $\widehat{d}'_r$, the fraction of recombination events that lead to uniquely observable recombinants. Recalling Equation 4.6, a method-of-moments estimator for the population recombination rate is then

$$\widehat{\rho} = \widehat{R}_{\min}/(L\,\widehat{d}'_r a_n) \tag{4.10}$$

Other Approaches for Narrow Intervals

An alternative method-of-moments approach to estimating ρ was suggested by Hudson (1987), who noted that the variance of pairwise measures of neutral sequence divergence is expected to decline with increasing levels of recombination. (With strong linkage disequilibrium, some random pairs of haplotype blocks will be identical over all polymorphic sites, while others will differ at all such sites). This approach requires an estimate of the average number of nucleotide differences between random sequences of length L, $\Theta_\pi = \theta_\pi L$, as well as one other summary statistic, the observed variance of pairwise divergence,

$$\widehat{\sigma}^2_k = \frac{2}{n(n-1)} \sum_{\substack{i=1 \\ j>i}}^{n} (k_{ij} - \Theta_\pi)^2 \tag{4.11}$$

where k_{ij} is the number of sites at which sequences i and j differ, and n is the number of chromosomes scored in the sample. Wakeley's (1997) Equation 15 allows one to

estimate ρL as a function of Θ_π, $\widehat{\sigma}_k^2$, and n, and simple division by the length of the sequence (L) then yields $\widehat{\rho}$.

Fuller use of the information in sample data can be achieved by considering the probabilities of various sample counts of the four gametic types at two loci or nucleotide sites (i.e., **AB**, **Ab**, **aB**, and **ab**) assumed to be biallelic, neutral, and in drift-mutation-recombination equilibrium (Hudson 2001). For any hypothetical combination of the parameters θ and ρL, one may compute the probability of the observed data for each pairwise combination of markers (Golding 1984; Ethier and Griffiths 1990), although obtaining exact probabilities of two-locus sampling configurations is mathematically challenging and for large sample sizes approximations must often be obtained by computer simulation (but see Jenkins and Song 2009). Further simplification is made possible by obtaining probabilities of sampling configurations conditional on two alleles actually segregating at both sites, as this eliminates the dependence on θ, provided the latter is small enough to ignore parallel segregating mutations (Hudson 2001). One can then combine the likelihood estimates with respect to ρ over all nonoverlapping pairs of linked segregating sites to obtain a global estimate of ρ (Hudson 2001). Again because the data are not entirely independent, this **composite likelihood approach** is just an approximation to a full ML analysis, and the confidence limits for the resultant estimates can only be achieved by computer simulations. McVean et al. (2002) extended this approach to allow for parallel mutations, which in species with high mutation rates can lead to the false appearance of recombination under the usual assumptions of the four-gamete test.

The efficiency of all of the above methods can be questioned in the sense that they use summary statistics that do not necessarily make full use of all of the information in the sample. Most notably, they do not account for the genealogical relationships among the sampled haplotypes. To this end, several more elaborate ML approaches and their Bayesian extensions go well beyond the method of Hudson (2001) (e.g., Kuhner et al. 2000; Nielsen 2000; Fearnhead and Donnelly 2001). As the number of genealogies consistent with any given set of mutational and recombinational parameters is enormous, exact solutions are not possible with these computationally intensive approximations. Moreover, although one would expect estimates derived in an explicit likelihood framework to perform better than the types of *ad hoc* procedures outlined above, it remains unclear whether that is the case for the sample sizes (n and L) that have been typically applied to date, as all existing estimators appear to be biased, have very large sampling variances, and rely on the assumption of an equilibrium population (Wall 2000).

Large-scale Analysis

The methods outlined in the preceding paragraphs were developed largely for analyzing sequences at the level of gene-sized fragments. However, with the sequencing of entire genomes of multiple individuals now becoming routine, entire genomic profiles of LD can be obtained. One limitation of this new technology is that sequencing read lengths remain small (often on the order of 100 bp), so that unlike the situation when individual alleles are cloned and sequenced, the phases of haplotypes are not

certain for double heterozygotes at distant pairs of sites. However, unambiguous haplotypes can still be inferred from information contained within singly heterozygous individuals, with the resultant frequency estimates enabling one to compute the full slate of LD statistics.

One approach to estimating ρ from whole-genome sequencing relies on data from just a single individual (Lynch 2008a). This ML method estimates the correlation Δ of "zygosity" (heterozygosity and homozygosity) of pairs of sites separated by specific distances (d) across the genome to obtain measures of disequilibrium that are nearly unbiased with minimal sampling variance. Spatial correlations of heterozygosity arise because recombination causes variation in coalesence times among chromosomal regions. In effect, this leads to clustering of heterozygous sites in long stretches of DNA that by chance have experienced little recombination and have long coalescence times. For any distance d between sites, Δ_d is defined as the deviation of the frequency of pairs of nucleotide sites with mixed zygosities from the random expectation

$$\Delta_d = 1 - \frac{H_1(d)}{2\pi(1-\pi)} \tag{4.12}$$

with $H_1(d)$ denoting the fraction of pairs of sites containing one heterozygote and one homozygote, and $2\pi(1-\pi)$ being the expected fraction of such mixed pairs under a random distribution given an average level of heterozygosity π.

For the situation in which the genome-wide patterns of variation are largely driven by mutation, recombination, and genetic drift, and the population is in equilibrium, using expressions from Ohta and Kimura (1969) for the two-allele model, it can be shown that

$$E(\Delta_d) \simeq \frac{\theta(1+2\theta)(18+\rho_d)}{2(1+\theta)A} \tag{4.13a}$$

where

$$A = 9 + 6.5\rho_d + 0.5\rho_d^2 + 19\theta\rho_d + 12\theta^2\rho_d + \theta\rho_d^2 + 54\theta + 80\theta^2 + 32\theta^3. \tag{4.13b}$$

(Lynch et al., in prep.) Note that as $\rho_d \to 0$, $E(\Delta_L) \to \theta(1+2\theta)/(1+7\theta)$, which is closely approximated by θ when $\theta \ll 1$ (which, as noted above, is generally the case). As $\rho_d \to \infty$, $E(\Delta_d) \to \theta(1+\theta)/\rho_d \simeq \theta/\rho$. Thus, given an estimate of θ, with estimates of average Δ_d for neutral sites separated by $d = 1$, 2, 3, etc., sites, each based on thousands to millions of pairs of sites, the decline in Δ_d with d can be used to infer ρ.

Another potentially powerful method for estimating ρ with population genomic data takes advantage of the standardized linkage disequilibrium r_D^2 introduced in Chapter 2. For neutral sites in drift-mutation equilibrium, Equation 2.29a gives a full expression for r_D^2 in terms of θ and ρ_d. However, provided $\theta \ll 1$ (which is always the case) and $\rho_d \gg \theta$ (which, as shown below, is generally the case for physically distant sites), Equation 2.29b simplifies to

$$r_D^2 \simeq \frac{10+\rho_d}{(11+\rho_d)(2+\rho_d)} \simeq \frac{1}{2+\rho_d} \tag{4.14}$$

The simplification to the right (Hill 1975; McVean 2002), which causes no more than 10% bias in estimating ρ_d, is often relied on in the literature (Hayes et al.

2003; Tenesa et al. 2007). We note in Example 2.7 that another commonly used approximation, $r_D^2 \simeq 1/(1+\rho_d,)$ has a more restricted meaning that limits its use with molecular data. As the sampling variance for r_D^2 for single pairs of polymorphic sites is generally very high, the usual strategy is to procure a large number of estimates for different pairs of informative markers separated by a certain window of physical distance, and then to pool these into a single estimate for that distance. Subtracting an expected contribution $1/n$ to r_D^2 resulting from finite sample size (Weir and Hill 1980) and rearranging Equation 4.14, leads to the estimator for sites separated by distance d,

$$\widehat{\rho} = \frac{1}{d}\left(\frac{1}{\widehat{r_D^2} - (1/n)} - 2\right) \tag{4.15}$$

A significant problem, often unappreciated, is that estimates of r_D^2 are often substantially biased if sample sizes are small or allele frequencies are extreme.

Before proceeding, it will be useful to consider the specific mechanics that cause recombination between nucleotide sites. Although it is often assumed that the recombination rate is simply equal to the crossover rate between sites, this is generally not true for closely-spaced sites. Recombination events nearly always involve heteroduplex formations between homologous chromosomes, i.e., the temporary physical annealing of homologous regions of complementary strands (usually no more than a few hundred base pairs). When such regions contain heterozygous sites, the nonmatching sites have to be resolved by gene conversion. Inclusion of this matter in the interpretation of the recombination rate is essential because although all recombination events result in gene conversion, not all gene conversion events are accompanied by crossovers. Because gene-conversion tracts are relatively short, when sites are far apart, most recombination events result from crossing over, but when sites are close together, recombination mostly results from the conversion of single sites.

To understand this in a more quantitative way, let c be the total rate of initiation of recombination events per nucleotide site (with or without crossing over), d be the number of sites separating the two focal positions (with $d = 1$ for adjacent sites), and x be the fraction of recombination events accompanied by crossing over. Using Haldane's (1919) mapping function, which assumes random and independent recombination at all sites, the crossover rate can be represented as $0.5(1 - e^{-2cxd})$, which is $\simeq cxd$ for $cxd \ll 1$, and asymptotically approaches 0.5 for large cd. In the following, we assume distances between sites that are small enough that $r_x \simeq cxd$. As noted by Andolfatto and Nordborg (1998), a gene conversion event has consequences equivalent to a crossover if the conversion tract is restricted to a single site. Under the assumption of an exponential distribution of tract lengths with mean length T (in bp), the total conversion rate per site is $(1-x)cT(1-e^{-d/T})$ (Langley et al. 2000; Frisse et al. 2001; Lynch et al. in prep.). The total recombination rate between sites separated by distance d is then

$$c_d \simeq c[xd + (1-x)T(1-e^{-d/T})] \tag{4.16a}$$

For $d \ll T$, most conversion events cover both sites, and

$$c_d \simeq cd \tag{4.16b}$$

whereas for $d \gg T$,

$$c_d \simeq cdx \tag{4.16c}$$

These results show that the simple division of an estimate of ρ_d by d to obtain an estimate of the per-site parameter ρ may yield rather different answers depending on the distance between sites, and that even at large distances ρ specifically measures the population crossing over rate between sites.

Empirical Observations

As in the case of estimates of $4N_eu$, all estimates of the per-site value of $\rho = 4N_ec$ are much smaller than 1.0 (Table 4.1). Indeed, all estimates are < 0.1, with many falling below 0.01, providing strong support for the idea that random genetic drift is a much more powerful force than recombination at the level of individual nucleotide sites. Moreover, by dividing estimates of ρ by parallel estimates of θ, the effective population size cancels out, yielding an estimate of the relative power of recombination and mutation at the nucleotide level (c/u). All such estimates are smaller than 5.0, and nearly half are smaller than 1.0, implying that the power of recombination between adjacent sites is generally of the same order of magnitude or smaller than the power of mutation (Table 4.1). For *Drosophila,* the average estimate of $c/u \simeq 2.7$, whereas for humans, it is ~ 0.8. Average c/u for fourteen land plants is 1.1 (SD = 1.2), although this may somewhat underestimate the average for purely outcrossing species because several of the species included in the survey (e.g., *Arabidopsis* and *Oryza*) are predominantly self-fertilizing, which reduces the effective amount of recombination (Hagenblad and Nordborg 2002).

Remarkably, even though prokaryotes do not engage in meiosis, estimates of c/u for such species are generally of the same order of magnitude as those for eukaryotes (Lynch 2007). This suggests, that relative to the background rate of mutation, recombination at the nucleotide level is not exceptionally low in prokaryotes, although the downward bias in estimates of θ for this group (noted above), may lead to inflated estimates of c/u.

Table 4.1. Estimates of the per-site population recombination rate ($\rho = 4N_ec$) and the ratio of the per-site recombination and mutation rates per nucleotide site (c/u, obtained by dividing estimates of ρ by estimates of $\theta = 4N_eu$). All estimates are derived from population surveys of nucleotide variation at silent sites in protein-coding genes.

Species	ρ	c/u	References
Animals:			
Drosophila melanogaster	0.05846	3.545	Hey and Wakeley 1997 Andolfatto and Przeworski 2000
Drosophila pseudoobscura	0.08655	1.360	Hey and Wakeley 1997
Drosophila simulans	0.09720	3.306	Andolfatto and Przeworski 2000
Homo sapiens	0.00060	0.770	Frisse et al. 2001; Ptak et al. 2004 Lefebvre and Labuda 2008
Land plants:			
Arabidopsis thaliana	0.00160	0.193	Kim et al. 2007

Brassica nigra	0.00602	0.330	Lagercrantz et al. 2002
Cryptomeria japonica	0.00046	0.118	Fujimoto et al. 2008
Helianthus annuus	0.05280	4.100	Liu and Burke 2006
Hordeum vulgare	0.00080	1.417	Morrell et al. 2006
Oryza rufipogon	0.00003	0.006	Mather et al. 2007
Oryza sativa	0.00004	0.021	Mather et al. 2007
Persea americana	0.00338	0.582	Chen et al. 2008
Pinus sylvestris	0.01452	2.855	Pyhäjärvi et al. 2007
Pinus taeda	0.00175	0.266	Brown et al. 2004
Solanum chilense	0.02380	1.122	Arunyawat et al. 2007
Solanum peruvianum	0.03480	1.392	Arunyawat et al. 2007
Sorghum bicolor	0.00041	0.130	Hamblin et al. 2005
Zea mays	0.02840	2.176	Tenaillon et al. 2004

EFFECTIVE POPULATION SIZE

Although the theory outlined in Chapter 3 suggests numerous ways in which the effective size of a population might be estimated from demographic data, such information is often difficult to come by, except in carefully controlled breeding populations. Moreover, estimates of N_e based on demography alone generally do not incorporate the long-term effects of selection on linked chromosomal regions, certainly not selective sweeps or background selection (Chapter 8). Nevertheless, given the effects that drift has on the temporal dynamics of neutral variation, there are a number of ways in which observations on the latter features can be used to indirectly infer N_e. From the standpoint of natural populations, two approaches harbor the most promise – monitoring temporal changes in putatively neutral allele frequencies, and ascertaining genome-wide patterns of LD, in both cases back-calculating the value of N_e that best explains the data (reviewed by Wang 2005).

Temporal Change in Allele Frequencies

Consider a single nucleotide polymorphism (**SNP**) sampled on two occasions separated by t generations, with initial frequency p_0, and recall from Chapter 2 that the expected variance in allele-frequency change after t generations is $p_0(1-p_0)(1-e^{-t/(2N_e)}) \simeq p_0(1-p_0)t/(2N_e)$ for small $t/(2N_e)$. This represents only the true population variance (the evolutionary variance in the preceding parlance), to which the sampling variance associated with *observed* allele-frequency estimates must be added. Summing these two sources of stochasticity yields an overall estimate of the expected variance of allele-frequency change of $p_0(1-p_0)[t/(2N_e)+1/(2n_0)+1/(2n_1)]$, where n_0 and n_1 denote the number of individuals (assumed to be diploid) genotyped in the two generations. Letting $\widehat{p}_0$ and $\widehat{p}_1$ be the estimated allele frequencies in the two generations, the expected variance in allele-frequency change across generations can also be written as $E[(\widehat{p}_1-\widehat{p}_0)^2]$ because $E(\widehat{p}_1-\widehat{p}_0)=0$ under neutrality.

Krimbas and Tsakas (1971) suggested that by equating these two quantities and rearranging, the effective population size can be estimated from observations over

two consecutive generations

$$\widehat{N}_e = \frac{1}{2\widehat{F}_1 - (1/n_0) - (1/n_1)} \tag{4.17a}$$

where

$$\widehat{F}_1 = \frac{(\widehat{p}_0 - \widehat{p}_1)^2}{\widehat{p}_0(1-\widehat{p}_0)} \tag{4.17b}$$

is a measure of the standardized variance of allele-frequency change. Provided $t/(2N_e) \ll 1$, the same expression applies when samples are made t generations apart, if t is substituted for one in the numerator of Equation 4.17a. (Note that the definition of F_1 is identical in form to the population-subdivision statistic F_{ST}, presented as Equation 2.42, except that the latter is concerned with spatial rather than temporal variation).

Despite their intuitive nature, Equations 4.17a,b yield biased estimates because the contributions of the sampling variance (and in some cases, covariance) of allele frequencies to F_1 are not fully accounted for (Pamilo and Varvio-Aho 1980; Nei and Tajima 1981b; Pollak 1983; Tajima and Nei 1984; Waples 1989). Additional limitations are that $\widehat{F}_1$ is undefined if $\widehat{p}_0 = 0$, and that Equations 4.17a,b do not immediately allow for the incorporation of multiple alleles ($k > 2$). An alternative estimator that deals with these problems is

$$\widehat{N}_e = \frac{t-2}{2\widehat{F} - (1/n_0) - (1/n_1)} \tag{4.18a}$$

where $\widehat{F}$ is calculated by either

$$\widehat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} \frac{(\widehat{p}_{0i} - \widehat{p}_{1i})^2}{[(\widehat{p}_{0i} + \widehat{p}_{1i})/2] - \widehat{p}_{0i}\widehat{p}_{1i}} \tag{4.18b}$$

(Nei and Tajima 1981b), or

$$\widehat{F}_3 = \frac{1}{k}\sum_{i=1}^{k} \frac{(\widehat{p}_{0i} - \widehat{p}_{1i})^2}{(\widehat{p}_{0i} + \widehat{p}_{1i})/2} \tag{4.18c}$$

(Pollak 1983). The details leading up to these alternative expressions can be found in the primary references, but it is notable that because $(\widehat{p}_{0i} + \widehat{p}_{1i})/2$ is generally much larger than $\widehat{p}_{0i}\widehat{p}_{1i}$, both estimators usually lead to very similar results (Waples 1989). One drawback of Equation 4.18a is that it requires an interval of at least three generations.

More refined measures of F can be obtained by averaging estimates of F_1, F_2, or F_3 over multiple loci, and Pollak (1983) derived a generalized estimator that allows for sampling across more than a single time interval. All of these approaches assume that the sampling of individuals at the beginning of an interval has no effect on the allele-frequency variance, which is reasonable when samples constitute a minor fraction of the population or are taken in a nondestructive manner or following reproduction. An additional concern is the sampling scheme for allele frequencies, which is straight-forward in a synchronized population with discrete

generations, but potentially problematical in species with overlapping generations. In the latter case, the contributions of sampled individuals to the overall allele-frequency estimates need to be weighted by the reproductive values of various age classes (Waples and Yokota 2007), a difficult enterprise with species with poorly understood life histories. Attention to these issues is provided in Nei and Tajima (1981b) and Waples (1989).

Regardless of the method used, estimates of N_e derived by these method-of-moment estimators generally have substantial sampling variances, and negative estimates of N_e are even possible. Clearly, if $t/(2N_e) \ll 1/(2n_0) + 1/(2n_1)$, observed fluctuations in allele frequencies will be largely a consequence of sampling error, so the utility of the overall approach becomes diminishingly small in populations with large effective sizes. Assuming equal sample sizes for each locus, the sampling variance of $\widehat{N}_e$ is

$$\mathrm{Var}(\widehat{N_e}) \simeq \left(\frac{8N_e^4}{t^2 M}\right)\left(\frac{1}{4N_e^2} + \frac{1}{N_e t\tilde{n}} + \frac{1}{t^2\tilde{n}^2}\right) \tag{4.19}$$

where M denotes the number of independent allelic comparisons (approximated by the sum of $k-1$ over all loci) and $\tilde{n}$ is the harmonic mean of the sample sizes in the two generations (Pollak 1983). In general, M, t, and $\tilde{n}$ will be under the control of the investigator, so the form of Equation 4.19 provides a useful basis for designing an optimal sampling strategy. For example, a doubling of M will reduce the sampling variance by one half, whereas a doubling of the sampling interval (t), which may often be less costly, has a much greater effect.

The sampling distribution of $M\widehat{F}/E(F)$ is expected to be approximately χ^2 in form, with M degrees of freedom (Lewontin and Krakauer 1973; Nei and Tajima 1981b), and this fact can be used to construct confidence intervals for N_e by substituting the critical χ^2 values for $\widehat{F}$ into Equation 4.18a, e.g., using the values of F at the 2.5 and 97.5% cumulative probability levels to yield 95% confidence limits. However, using computer simulations, Goldringer and Bataillon (2004) found that the χ^2 assumption can be significantly violated when there is a minor allele (with frequency < 0.1), a large number of alleles (as with microsatellites), or the number of generations between sampling times is large. In Chapter 9, the issue of temporal change in allele frequencies will be revisited from a different perspective – testing the hypothesis that an observed magnitude of change is inconsistent with random genetic drift for an assumed value of N_e, or equivalently estimating the largest value of N_e that is consistent with the observed change being entirely due to drift.

Finally, it is worth noting that because of their simple heuristic interpretation, method-of-moments estimators, like those just noted, are highly popular approaches for estimating population parameters. However, by relying on a single summary statistic, such methods do not fully utilize the information in a set of samples. A more powerful approach to estimating N_e from sequential samples involves the use of ML procedures (and their Bayesian extensions) to yield estimates that best explain the entire distribution of observed allele frequencies conditional on sample sizes (Williamson and Slatkin 1999; Anderson et al. 2000; Berthier et al. 2002). These methods are highly demanding computationally, to a degree that increases with N_e, although Wang (2001), Beaumont (2003), Tallmon et al. (2004), Anderson (2005), and Bollback et al. (2008) present computationally efficient approximations.

Single-sample Estimators

Because of the practical difficulties in obtaining temporal sequences of samples, especially in species such as vertebrates and land plants with long generation times, a number of methods have been developed for estimating N_e from the information contained in just a single sample. Only a brief overview of such methods will be provided here. One of the most commonly applied single-sample estimators is the LD method, already outlined above. Under the assumption of drift-mutation-recombination equilibrium, estimates of $\rho = 4N_e c$ can be obtained, so if the recombination rate between the loci under consideration is known, then $\widehat{\rho}/(4c)$ will provide an estimate of N_e (Hill 1981). Likewise, if the mutation rate per nucleotide site (per generation) is known, any estimate of $\theta = 4N_e u$ can converted to an estimate of long-term N_e using $\widehat{\theta}/(4u)$. Again, being based on simple summary statistics, these estimators do not utilize all of the information inherent in a sample of alleles, and hence are not likely to provide the most efficient estimates of N_e. As an alternative, highly computationally intense coalescent sampling methods have been developed to estimate various population-genetic parameters, including N_e, mutation, and recombination rates, and other demographic parameters (e.g., population growth rates and degree of subdivision), using the genealogical information inherent in samples of population sequences (Kuhner 2008).

A second approach, applicable only to randomly mating species, relies on observed amounts of excess heterozygosity relative to Hardy-Weinberg expectations. The basis of this procedure is the random deviation in allele frequency that develops among the two sexes as a consequence of stochastic sampling of gametes in the preceding generation. In effect, the two sexes are being viewed as two random samples of gametes here – the smaller the value of N_e, the larger the expected deviation between the sexes (Robertson 1965; Pudovkin et al. 1996; Luikart and Cornuet 1999). Such variation among the sexes causes excess heterozygosity in the progeny generation by elevating the likelihood that each sex will contribute an alternate allele to offspring.

A third single-sample method attempts to estimate the fraction of pairs of randomly sampled offspring in a population that are full- or half-sibs (Wang 2009). This method relies on statistical procedures for deriving estimates of relatedness with molecular markers. As in the case of the heterozygosity-excess approach, information on a large number of informative markers is required, and a random sampling scheme is essential. These last two approaches are restricted to very small population sizes (on the order of 100 reproductive adults or fewer), as such conditions are required to generate detectable deviations from Hardy-Weinberg expectations and detectable numbers of sib pairs.

Empirical Observations

In Chapter 3, we found that the numerous demographic factors influencing the effective size of a population almost always do so in a downward direction. Applications of the methods outlined above provide some indication as to the magnitude of this reduction relative to the actual size of a population (N). Because the temporal-

fluctuation method requires a small enough N_e to yield meaningful results on a reasonable time scale, not surprisingly, almost all estimates using this technique derive from large-bodied vertebrate species. In a survey of studies on mostly low-fecundity species, Frankham (1995) found an average N_e/N of ~ 0.11, whereas a subsequent study with a much larger sample obtained an average of 0.14 (Palstra and Ruzzante 2008).

It is likely that the $\sim 90\%$ reduction in N_e suggested by these studies is a considerable *underestimate* of the situation for many nonvertebrate species and even many vertebrates. For example, as noted in Chapter 3, high-fecundity fish in spatially variable environments appear to have $N_e/N < 0.001$ (Hedrick 2005). In addition, many unicellular species have conspicuous phases of asexual reproduction that can encourage the rapid proliferation of a small number of clones, generating N_e/N ratios much lower than 0.001. Strongly inbreeding species (e.g., self-fertilizing plants) may also approach such extremes. Finally, one of the major short-comings of the temporal-fluctuation approach to estimating N_e may be its tendency to overlook rare, but quantitatively significant, phases in which genomic regions are exposed to strong selective sweeps at linked loci (Chapter 8).

Example 4.3. Hill (1981) noted that estimates of N_e based on the amount of standing LD between tightly linked markers are more a function of the long-term population history while LD measures between more loosely-linked markers are more reflective of recent history. Hayes et al. (2003) seized upon this observation to suggest that by using estimates of $\rho = 4N_e c$ for different values of c (i.e., known genetic-map distances between sites), one could in effect estimate the effective population sizes at different times in the past. In particular, they showed for a model of linear population change (growth or decline) that examining LD between markers with recombination frequency c estimates N_e at roughly $1/(2c)$ generations in the past.

Using this approach, Tenesa et al. (2007) scored roughly one million SNPs to examine LD at various distances for four different human populations. Estimates of historical N_e were obtained for each autosome, and Figure 4.3a shows the result for a Utah population of European ancestry. For given slices of time, the various points indicate the 22 separate estimates based on each autosome. Note both the consistency of estimates over autosomes and the very recent expansion of population size. Similar studies in humans were performed by Sved et al. (2008) and McEvoy et al. (2011). Hayes et al. (2003) and Flury et al. (2010) applied this approach to modern dairy cattle, showing in this case that N_e has dramatically declined from historical values, presumably reflecting the bottlenecking effects of selection for improvement (Figure 4.3b).

MUTATION RATE

The long-term evolution of complex traits ultimately depends on the input of new variation via mutation, which is a function of the rate at which new mutations arise at the DNA level and their influence at the phenotypic level, the combined effects defining the overall rate of polygenic mutation (LW Chapter 12). Here, we continue to focus specifically on the DNA-sequence level, with u being defined as the rate of mutation per nucleotide site per generation. Because mutations arise at an extremely low rate at most nucleotide sites, the direct estimation of u is formidably challenging, with most approaches relying on procedures that enrich the pool of experimentally derived mutations in an effectively neutral fashion (so that selection does not bias the outcome). Here, we review the two most commonly used methods of enrichment: 1) long-term genome-wide accumulation of mutations in isolated lineages with tiny effective population sizes; and 2) short-term isolation of conspicuous mutants at single marker loci from large populations raised on selective media.

Divergence Analysis

The most conceptually simple approach, frequently applied to multicellular organisms with fairly long generation times, is to perform a mutation-accumulation experiment (LW Chapter 12), whereby a set of initially genetically identical (and usually homozygous, if not clonal) lines are passed through repeated population bottlenecks. For example, with the self-fertilizing nematode *Caenorhabditis elegans* and plant *Arabidopsis thaliana,* an ancestral line can be repeatedly selfed to ensure homozygosity, with the progeny of one parent being used to synchronously initiate a set of parallel lines, each to be subsequently maintained by single-progeny descent. With each line having an effective population size of just one individual under this design, essentially all mutations that do not cause lethality or complete sterility (the vast majority of mutations) will accumulate independently at a rate u, in accordance with the neutral theory (Chapter 2). Under self-fertilization, newly arisen mutations are fixed or lost in just two generations on average, so after several dozens to hundreds of generations of mutation accumulation, nearly all fixed mutations can be detected as homozygotes by sequencing a subset of lines. Typically, nearly all lines will be identical at individual nucleotide sites (reflecting the ancestral state), with mutations appearing as single-line outliers.

Letting n denote the number of sites surveyed, L the number of lines, T the average number of generations per line, and m the number of observed mutations, the mutation rate per site is estimated as

$$\widehat{u} = m/(nLT) \tag{4.19a}$$

with sampling variance of

$$\sigma^2(\widehat{u}) \simeq \widehat{u}/(nLT) \tag{4.19b}$$

The latter expression implies a coefficient of sampling variation for $\widehat{u}$ of $(unLT)^{-1/2}$, which is the inverse of the square root of the expected number of observed mutations in the assay.

Example 4.1. A commonly used variant of the laboratory mutation-accumulation experiment for estimating mutation rates exploits the information inherent in natural populations, relying on presumptively neutral sequences from isolated but closely related species. Recall from Chapter 2 that the long-term rate of nucleotide substitution at neutral sites is equal to the mutation rate regardless of N_e, and from above that the average nucleotide divergence of random alleles within a species has expected value $4N_eu$. Thus, for two sister taxa that became isolated t generations in the past, the expected divergence of orthologous neutral sequences (number of substitutions per site) is $d = 2tu + 4N_eu$, assuming equal N_e in both taxa. At $t = 0$, $d = 4N_eu$ (the average divergence of randomly sampled alleles in the ancestral population), whereas as $t \to \infty$, $d \simeq 2tu$ (a widely used approximation in applications of molecular clocks for dating evolutionary events). Rearranging, and letting $\overline{\theta}_H$ be the estimate of the average within-species nucleotide diversity at silent sites $4N_eu$, we obtain an estimator for the mutation rate, $\widehat{u} = (\widehat{d} - \overline{\theta}_H)/(2t)$.

Nachman and Crowell (2000) used this approach to obtain an estimate of the mutation rate for humans from sequences of 12 unexpressed pseudogenes in human and chimpanzee. Because they are nontranscribed, such stretches of DNA are expected to fulfill the assumptions of neutrality. The average number of substitutions per site separating the two species was $\widehat{d} = 0.0133$. A broad geographic survey of within-species variation in 49 noncoding (and presumably largely neutral) regions yielded estimates of 0.00087 for human and 0.00134 for chimpanzee (Yu et al. 2003), implying $\overline{\theta}_H = 0.00110$. Nachman and Crowell assumed a divergence time of 5 million years, and an average generation time of 20 years, yielding $t \simeq 250,000$ generations. Substitution into the preceding expression then gives an estimated mutation rate of 2.44×10^{-8} per site per generation for base-substitution mutations, which strictly speaking is an average over the chimpanzee and human lineages.

Short-term Enrichment

The preceding approach employs a strategy of augmenting the pool of observable mutations by passing lines through a large number of generations. The advantage of such a protocol is that mutations are equally enriched throughout the genome, minimizing the chances that the mutational profile will be biased by observations at any particular target locus. The disadvantage of this procedure is that an enormous number of sites (typically many tens of millions) need to be searched to obtain just a few dozen mutations.

An alternative approach, widely applied to microbial cultures, focuses on reporter constructs (specific marker loci at which at least a subset of mutations causes obvious phenotypic changes). Here the emphasis is on the efficient screening of a very large pool of cells in a relatively short period of time for a small subset of newly arisen mutations, e.g., exponentially growing an initially nonmutant stock to a population size in excess of the reciprocal of the mutation rate (so there will clearly be more than one mutational event in the culture), and then isolating the subset of cells that have acquired a mutation at a locus that is nonessential in the background environment but permits subsequent growth on a selective medium (Luria and Delbrück 1943). From estimates of the total number of mutant and nonmu-

tant cells in the culture, it is then possible to determine the mutation rate per cell division.

Because mutant cells grow during culture expansion, the relationship between the number of mutant cells observed in a population and the actual number of mutational events that produced them is generally not one-to-one. Thus, the first challenge is to convert the observed number of mutant cells to the number of mutations leading to them (m). In addition, because not all mutations produce an observed phenotype, the second challenge is to determine the fraction of mutations that are detectable at the target locus (d). The true number of mutations is estimated by m/d. Finally, in order to determine the mutation rate per nucleotide site, one must know the mutational target size (n, in base pairs). Thus, for the marker approach to yield reliable estimates of u, a good deal of knowledge must exist on the molecular features of the target locus.

Several methods exist for estimating the number of unique mutational events from the observed numbers of mutant and nonmutant cells, with broad overviews being provided by Rosche and Foster (2000) and Angerer (2001a,b). Suppose a large series of replicate cultures is developed, and one then simply scores the fraction of cultures at the end point that are completely free of mutations (p_0). Assuming that the number of mutational events per culture is Poisson distributed with expectation m, the expected frequency of mutation-free cultures is then simply

$$E(p_0) = e^{-m} \tag{4.20}$$

Rearrangement leads to the estimator $\widehat{m} = -\ln(p_0)$, ignoring the sampling bias resulting from the error in estimating p_0. This approach works well when m is on the order of 0.5 to 2.5, but with more extreme values, p_0 will be close enough to 0.0 or 1.0 that meaningful estimates are not possible unless the number of cultures is enormous. A second disadvantage of this approach is its failure to use most of the information in the set of cultures, as the distribution of mutant numbers among replicate cultures is completely ignored. Full use of such information can be incorporated into a maximum-likelihood framework (e.g., Lea and Coulson 1949; Sarkar et al. 1992).

Example 4.3. An alternative approach to estimating the mutation rate in an exponentially growing culture is to consider the expected temporal dynamics of the frequency of mutant cells in the population. Letting f_0 be the initial frequency of mutations, r be the rate of exponential growth of the numbers of cells in the culture (assumed to be identical for cells that are mutant and nonmutant at the marker locus), and u_o be the rate of mutation to an observable phenotype per cell division, the expected frequency after t time units is

$$f_t = f_0 + (1 - f_0)(1 - e^{-u_o rt})$$

This follows from the fact that $e^{-u_o rt}$ is the probability that a descendant of a nonmutant cell has not acquired a detectable mutation after rt cell divisions. Note that if one starts with a mutation-free culture ($f_0 = 0$), and the cumulative probability of mutation ($\simeq u_o rt$) is $\ll 1$, the expected fraction of mutant cells will increase in an essentially linear fashion, at rate $u_o r$.

Because of the stochastic nature of mutations, results from single cultures are not terribly reliable with this approach. Thus, motivated by the original design of Luria and Delbrück (1943), most studies of microbial mutation grow a moderate number of initially mutation-free cultures up to an arbitrarily large population size, surveying the frequency of mutants at the end point of each culture. The simplest approach involves rearrangement of the preceding expression to yield the relevant point estimator of the mutation rate to observable phenotypes,

$$\widehat{u}_o = -\frac{\ln[(1-f_0)/(1-f_t)]}{rt}$$

Because $N_t = N_0 e^{rt}$ under exponential growth, where N_0 and N_t are the total numbers of cells in the culture at times 0 and t, so long as the observed mutant frequencies are < 0.1, so that $\ln(1-f) \simeq -f$, this expression further simplifies to

$$\widehat{u}_o \simeq \frac{f_t - f_0}{\ln(N_t/N_0)}$$

which is simply the rate of accumulation of observable mutations per cell division.

Drake (1991) has argued that this essentially deterministic view of the rate of increase of mutants is unlikely to hold very well until the culture has reached a large enough size to harbor at least some mutations, which is expected to take several generations. Taking the view that a reasonable benchmark is the point at which the culture is expected to contain a single mutant, which implies $u_o N = 1$, then one may take $f_0 = u_o$ and $N_0 = 1/u_o$ as an arbitrary starting point, which after substitution into the previous expression leads to

$$\widehat{u}_o \simeq \frac{f_t - u_o}{\ln(u_o N_t)}$$

Given just the total number of cells, N_t, and the frequency of mutants at the end point, f_t, this expression can be solved recursively to obtain the estimate $\widehat{u}_o$. When data are available from multiple cultures, f_t is generally taken to be the *median* frequency of mutants, as the mean can be strongly biased in the event the sample includes any "jackpot" cultures that happened to have acquired a mutation during an early cell division.

Conversion of the rate of origin of *observable* mutations, u_o, to an estimate of the mutation rate at the nucleotide level requires that the fraction of mutations that are detectable at the marker locus (d) be known. Many mutations have no phenotypic effects, e.g., because they arise at silent sites or at amino-acid replacement sites that have no substantive effect on the causal locus. To determine the fraction of undetectable mutations, a large number of independent mutant cells can be sequenced to ascertain the molecular basis of the changes at the target locus, and the degree to which these are concentrated at particular sites. Generally, because the mutation rate per nucleotide site is quite low, no more than a single change is found within a sequenced locus, so there is little ambiguity as to the identity of causal mutations.

For base-substitutional mutations, Drake (1991) made the following argument for obtaining an estimate of d. Assuming that all mutations causing premature translation termination (so-called nonsense mutations) cause functional changes that are detectable, then letting n_n denote the number of such mutations observed in the sample, the expected total number of base-substitutional mutations per sequence in the sample

(whether recorded as mutants or not) is $64n_n/3$. This follows from the fact that of the 64 possible triplet codons, three encode for chain termination (in most species), and assumes random mutation to all 64 codons. Thus, letting n_o denote the total number of observed base-substitutional mutations in the set of sampled sequences (missense and nonsense mutations), $\widehat{d} = 3n_o/(64n_n)$ provides an estimate of the fraction of base-substitutional mutations that are detectable (if all detected base-substitutional mutations were to termination codons, implying no effects of missense mutations, $n_n/n_o = 1$, and $\widehat{d} = 3/64$). If n is the length of the target sequence (in base pairs) over which mutations are detectable (generally assumed to be the length of the coding region, which could be an overestimate), an estimator for the base-substitutional rate per nucleotide site is then

$$\widehat{u} = \frac{\widehat{u}_o}{\widehat{d}n}$$

Empirical Observations

Although accurate estimates of the mutation rate are available for only a handful of species, some generalizations can be made. Estimated rates of base-substitutional mutation ($\times 10^{-9}$ per site per cell division) are on the order of 0.5 for reporter-construct studies and the complete sequencing of mutation-accumulation lines of the yeast *Saccharomyces cerevisiae* (Lynch 2006; Lang and Murray 2008; Lynch et al. 2008; Nishant et al. 2010). On a per-generation basis, they are ~ 5.6 and 5.4, respectively, for sequenced mutation-accumulation lines of the fly *Drosophila melanogaster* (Haag-Liautard et al. 2007; Schrider et al. 2013) and the nematode *Caenorhabditis elegans* (Denver et al. 2004, 2012), and 6.2 for the model plant *Arabidopsis thaliana* (Ossowski et al. 2010). The average base-substitutional mutation rate for ten prokaryotic species is $\sim 1.0 \times 10^{-9}$ per site per cell division (Lynch 2010). Taken together, these and additional data from other species imply a strong positive scaling of the mutation rate per generation with genome size (Figure 4.4). The bulk of small-scale mutations in genomes generally involve base-substitutions, with the ratio of insertion/deletion mutations to the former being in the range of 0.05 to 0.25 in yeast, *Drosophila,* and humans, but as high as ~ 1.0 in *C. elegans* (Lynch et al. 2008; Lynch 2009b).

-Insert Figure 4.4 Here-

Example 4.4. To indirectly estimate the human mutation rate, Kondrashov (2003) took advantage of records on genetic pathologies attributable to dominant mutations at known causal loci. The population frequency of genetic disorders (I, incidence) caused by dominant autosomal mutations provides a simple basis for estimating the mutation rate to defective alleles. This is because the expected frequency of a dominant

deleterious allele under selection-mutation balance is $p \simeq u/s$, where u is the mutation rate to defective alleles (per gene copy), and s is the selective disadvantage of affected (heterozygous) individuals (Equation 7.6b). For a severe disorder, the frequency of the deleterious allele will be so small that essentially all affected individuals are heterozygotes, implying an incidence of the disorder very close to $2p(1-p) \simeq 2p = 2u/s$. Thus, the mutation rate to dominant defective alleles can be estimated as $sI/2$. (For a dominant mutation that leads to complete loss of reproductive fitness, $s = 1$, and the incidence is simply equal to $2u$, as each functional parental allele has a probability u of mutating to a defective product).

The remaining challenge is to convert the total rate of observed mutations at a locus to the underlying rate at the level of individual nucleotide sites. This can be accomplished by employing a strategy similar in spirit to that advocated by Drake (1991). For each disorder in the survey of Kondrashov (2003), a large sample of affected individuals (whose parents were known to be nonmutant) had both of their alleles sequenced to identify the nature of the newly arisen, causal mutations. Assuming all insertion/deletion mutations were detectable, from the incidence of chain-terminating base-substitutional mutations, the total detectability of mutations could then be calculated, as outlined in Example 4.2. Although Kondrashov's (2003) survey involved 32 different genetic disorders (each determined by a unique locus), we will simply present the calculations for one such analysis, and conclude with a summary of all of the results.

Familial adenomatous polyposis is a genetic disorder known to be caused by dominant mutations in the adenomatous polyposis coli (APC) tumor-suppressor gene, arising at an estimated rate of $u_o = 7 \times 10^{-6}$ per gene copy per generation. Of the 799 mutations validated by sequencing and deemed to be causal, 202 involved nonsense base substitutions, with the remaining 597 being associated with major lesions, insertions, and deletions of various sorts. Assuming that the total number of base substitutional mutations (when extrapolated to unaffected mutants) is $202 \times (64/3)$, and that all insertions and deletions are detectable, the overall detectability is estimated as $799/[597 + (202 \cdot 64/3)] = 0.163$. From the pool of affected individuals subjected to sequencing, a fraction 0.325 exhibited no causal mutation (presumably because the mutation resided outside of the sequenced target exons, which summed to 4803 sites). The estimated total mutation rate at the locus is therefore $(7 \times 10^{-6}) \times 0.675/(4803 \times 0.163) = 6.0 \times 10^{-9}$ per site per generation, a fraction $1 - \{597/[597 + (202 \cdot 64/3)]\} = 0.878$ of which involves base-substitutional mutations.

When these approaches are extended to the remaining 31 loci, the estimated average total mutation rate to base-substitutional changes is 1.70×10^{-8} per site per generation, averaged over both sexes. A more recent estimate involving a larger number of loci underlying human genetic disorders and somewhat different assumptions yielded an estimate of 1.29×10^{-8} (Lynch 2009b). More recently, direct estimates of the human mutation rate have also been generated by whole-genome sequencing in known lines of descent. For example, from information on portions of Y chromosomes separated by 13 generations of paternal-line descent, Xue et al. (2009) obtained a base-substitutional mutation rate estimate of 1.73×10^{-8} after scaling across the sexes to account for the lower rate of mutation in females. Three additional studies, involving autosomal sequences of parent-offspring trios, all yield sex-averaged estimates close to 1.2×10^{-8} (Conrad et al. 2011; Campbell et al. 2012; Kong et al. 2012).

Taken together, these estimates point to a sex-averaged base-substitutional mutation rate of $\sim 1.4 \times 10^{-8}$ for humans. This estimate is significantly lower than the phylo-

genetic estimate reported in Example 4.1 (2.44×10^{-8}). A number of factors might account for the elevated rate based on interspecies divergence: an incorrect estimate of the time of divergence between the human and chimpanzee lineages; an incorrect estimate of the amount of heterozygosity at initial divergence; inaccurate estimates of average generation times since the time of divergence; an elevated rate of mutation in the chimpanzee lineage; a recent decline in the human mutation rate; the operation of some selection on the sites analyzed; etc. The main point here is that estimates of the mutation rate derived from phylogenetic data are subject to numerous sources of potential error, the magnitude of which is generally unknown (and in some cases unknowable).

Evolution of the Mutation Rate

What are the likely mechanisms driving the pattern in Figure 4.4? There is no evidence that genome size directly influences the mutation rate per nucleotide site. Rather the relationship between the two traits is likely to be an indirect consequence of some shared factor. One obvious distinction among the above-mentioned groups is that multicellular species experience multiple germline cell divisions per generation (Lynch 2010), e.g., ∼10 for *C. elegans,* 36 for *D. melanogaster,* 40 for *A. thaliana,* and 200 for *H. sapiens* (Drost and Lee 1995; Kimble and Ward 1998; Crow 2000; Lynch 2010), whereas there is one cell division per generation in unicellular species. If most mutations arise as replication errors, one would then expect the per-generation mutation rate to scale across yeast : *C. elegans* : *D. melanogaster* /*A. thaliana* : human in an approximately 1 : 10 : 38 : 200 ratio, whereas the per-generation mutation-rate scaling implied by the results given above is less extreme, approximately 1 : 8 : 13 : 34. Mutation rates of microsatellite loci, which mutate via changes in nucleotide-motif repeat numbers, are also magnified with the level of multicellularity, but the ratio of per-generation mutation rates is on the order of 1 : 50 : 13,400 for unicellular eukaryotes, invertebrates, and mammals (Seyfert et al. 2008), which is much more extreme than the scaling of germline-cell division number. Thus, it appears that additional factors, including those independent of replication, must be responsible for the pattern exhibited in Figure 4.4.

As in the case of all phenotypic traits, the rate of mutation is certainly subject to the forces of natural selection (Baer et al. 2007). However, selection on the mutation rate is unusual in that the phenotypic effects associated with a mutator or antimutator allele are generally only manifested indirectly through the mutational changes induced at other fitness-related loci. This raises the question as to whether mutation rates are typically held at optimum intermediate levels by stabilizing selection so as to maximize the long-term rate of adaptive evolution, or simply pushed to their physiologically defined lower limits. If replication-error rates are maintained at higher levels than can be explained by constraints on cellular processes, the next obvious question is why dramatically higher mutation rates would be selectively promoted in multicellular relative to unicellular species, despite the fact that most mutations are deleterious (LW Chapter 12). However, the most central difficulty

with arguments that invoke long-term benefits of elevated mutation rates is that high mutation rates are much more likely to evolve in predominantly asexual populations (the situation in many unicellular species, but not multicellular taxa), as an absence of recombination is essential if novel mutator alleles are to be pulled to fixation via linkage to induced beneficial mutations (Johnson 1999a; Sniegowski et al. 2000; Wilke et al. 2001; André et al. 2006; Denamur and Matic 2006). Yet, as noted above, it is in unicellular and predominantly asexual species where the lowest mutation rates are consistently observed.

Despite substantial theoretical research, it has proven quite difficult to avoid the conclusion that mutation rates are predominantly driven downwardly by the transient linkage of mutator alleles to their recurrent deleterious side effects (Sturtevant 1937; Leigh 1970, 1973; Johnson 1999b). Occasionally, a mutator allele may be brought to high-frequency by hitch-hiking with a tightly linked beneficial mutation (Clune et al. 2008; Desai and Fisher 2011), but such events are expected to be transient, as they are quickly followed by loss of the mutator phenotype by either recombinational decoupling or reversion of the mutation rate. To see why recurrent deleterious mutation imposes selection against mutator alleles, note that any allele that magnifies the mutation rate (hereafter, designated as a mutator allele) will necessarily generate statistical associations with defective germline mutations induced at linked and unlinked loci. The duration of such disequilibria will depend on the rate of recombination between the mutator and affected loci, but because new associations will arise recurrently each generation by mutation, an equilbrium background mutation load will eventually be reached, with alleles imposing higher mutation rates developing a higher associated deleterious load.

Consider a locus relevant to fitness that recombines at rate c with respect to the mutator locus. If, in the heterozygous state, the mutator induces mutations at the fitness locus at an elevated rate Δu per gene with a reduction in fitness equal to hs per induced mutation, the selective disadvantage of the mutator allele induced by linkage disequilibrium with this particular fitness locus is

$$s_d \simeq \frac{hs \cdot \Delta u}{1-(1-hs)(1-c)} \tag{4.21a}$$

assuming $\Delta u \ll hs$ (Kimura 1967; Dawson 1999). For unlinked loci ($c = 0.5$), as in freely recombining species, this expression reduces to

$$s_d \simeq \frac{2hs \cdot \Delta u}{1+hs} \tag{4.21b}$$

whereas in the absence of recombination ($c = 0.0$),

$$s_d = \Delta u \tag{4.21c}$$

Because known average deleterious fitness effects of mutations imply $hs \ll 0.1$ (LW Chapter 12), these results indicate that the strength of selection opposing the downward drive of mutation rate is much weaker in sexual than in asexual species.

It follows from Equation 4.21c that, provided the equilibrium load associated with selection-mutation balance is reached, the total magnitude of selection against a mutator allele in an asexual population is simply equal to the elevation in the

genome-wide deleterious mutation rate (ΔU, summed over all fitness-relevant loci), independent of the effects of the mutations. However, the total disadvantage of a mutator in a sexual species must take into consideration mutations arising both on the chromosome carrying the mutator and on all other unlinked loci, as only tightly linked loci remain in association with the mutator for more than a few generations. Assuming L chromosomes, each one Morgan in length (below), and a haploid genome-wide increase in the deleterious mutation rate of ΔU, after accounting for the spatial distribution of random mutations, the total induced selection coefficient against the mutator allele is found to be

$$s_{d,T} \simeq \frac{2hs \cdot \Delta U(L - 1 + \phi)}{L(1 + hs)} \tag{4.22a}$$

where

$$\phi = 1 + \ln\left(\frac{1 + hs - (1 - hs)e^{-1}}{2hs}\right) \tag{4.22b}$$

is the approximate elevation in the average induced fitness effect of mutations on the mutator-bearing chromosome relative to that on the other $L - 1$ unlinked chromosomes (Lynch 2008b). For $0.001 < hs < 0.1$, which fully covers the range of average mutational effects found in empirical studies (LW Chapter 12), ϕ is in the range of two to seven. Thus, the selective disadvantage of a mutator allele in a sexual species is close to twice the product of the heterozygous fitness effect of new mutations (hs) and the haploid genome-wide increase in the deleterious mutation rate (ΔU) unless the chromosome number is very small, and even then not likely to be much more than a few-fold higher. The factor by which $s_{d,T}$ exceeds $hs \cdot \Delta U/(1+hs)$ is equivalent to the average number of generations that an induced deleterious mutation remains associated with the mutator responsible for its origin (as can be seen from Equation 4.21b, this factor is two for unlinked loci).

Because single amino-acid substitutions in DNA-processing proteins may have arbitrarily small effects on the mutation rate, and because existing mutation rates are already so low that there is little further room for improvement (the maximum possible reduction being the mutation rate itself), these results imply that the long-term selective disadvantage of many mutator alleles may be sufficiently small (relative to the power of genetic drift) to render them immune to the eyes of natural selection (Chapter 7). Thus, because there is a substantial decline in N_e from microbes to small invertebrates to vertebrates and large land plants (Lynch 2007), it is plausible that the elevation of mutation rates in multicellular lineages is not an inevitable consequence of an inherent physiological limitation in such species, but a simple consequence of the diminished ability of natural selection to enhance the level of replication fidelity (Lynch 2011; Jain and Nagar 2013).

Several observations are consistent with this **drift-barrier hypothesis**. First, for the set of species with adequate data, there is an inverse relationship between the mutation rate per nucleotide site per generation (u) and N_e (Sung et al. 2012). Second, empirical observations on the molecular machinery involved in DNA replication and repair indicate that these processes are indeed more error-prone in taxa with higher overall mutation rates (Lynch 2008a,b; 2011). Third, u is also inversely proportional to the number of a functional genes in a genome (Drake et al. 1998; Massey 2008; Ness et al. 2012; Sung et al. 2012). This relationship is expected

because, as noted above, selection operates on the total rate of deleterious-mutation production across the genome, which increases with the mutational target size. Finally, long-term laboratory-evolution experiments starting with mutator strains of microbes often reveal a gradual reduction in the mutation rate resulting from the spontaneous accumulation of changes at diverse genomic locations (Tröbner and Piechocki 1984; Notley-McRobb et al. 2002; Herr et al. 2011; Weigloss et al. 2013; Williams et al. 2013). Such observations clearly demonstrate that, even in microbes where the efficiency of selection is expected to be strong, the loci underlying replication fidelity have not been driven to a point where further improvement is no longer possible.

The central point here is that one of the primary determinants of the evolutionary features of a population, the mutation rate itself, is subject to substantial evolutionary modification, with the effective population size and number genomic sites of functional significance dictating the degree to which selection can reduce the replication-error rate. The pattern in Figure 4.4 emerges not because there is a direct causal connection between total genome size and the mutation rate, but simply because eukaryotic genomes become bloated in size via the accumulation of substantial noncoding DNA with increasing magnitudes of genetic drift (Lynch 2007).

RECOMBINATION RATE

Although it is extraordinarily difficult to estimate recombination rates at specific nucleotide sites, some compelling general statements can be made about *average* levels of recombination over entire genomes. Such information derives from high-density genetic maps constructed from observed rates of meiotic crossing-over between molecular markers, now available for hundreds of eukaryotes thanks to the widespread availability of highly variable markers such as microsatellites. Genetic maps are based on mapping functions that attempt to convert observed recombination frequencies into the expected numbers of crossover events between pairs of markers (LW Chapter 14). Strictly speaking, such maps measure the frequency of crossover events, and generally do not include the added contributions of gene conversion, which can cause the recombination rate between very closely spaced sites to exceed by several fold the expectation based on distant markers that are predominantly rearranged by crossovers (Equations 4.16a-c). Chromosome lengths are generally reported in units of **Morgans** (the average number of crossovers per chromosome), with the sum of these lengths over all chromosomes giving the total map length.

Although eukaryotic genome sizes (total numbers of nucleotides) vary by four orders of magnitude, the range of variation in genetic-map lengths among species is only about ten-fold, with the averages for various phylogenetic groups deviating by only five-fold (Table 4.3). A simple physical constraint explains such behavior. During meiosis, there are typically no more than two crossover events per chromosome (one per arm), so that average chromosome lengths are generally on the order of one Morgan, regardless of chromosome size. Thus, because phylogenetic increases

in genome size are generally associated with increases in chromosome size rather than chromosome number (Table 4.3), there is little variation in the total amount of meiotic crossing over per genome across a vast swath of life.

These observations lead to a simple structural model for the average recombination rate per physical distance across a genome ($\bar{c}$). Letting G be the total number of bases per haploid genome, and N be the haploid number of chromosomes per genome, G/N is the mean physical length of chromosomes. Letting x be the average number of crossovers (Morgans) per chromosome per meiosis, then $\bar{c} \simeq xN/G$, assuming that x is independent of chromosome size. If this model is correct, a regression of $\bar{c}$ on G on a log scale should have a slope not significantly different from -1.0, with the vertical distribution (residual deviations) around the regression line being defined largely by variation in xN (the total number of crossovers per genome). The data closely adhere to this predicted pattern, with the smallest genomes of microbial eukaryotes having recombination rates per physical distance that are $\sim 1000\times$ greater than those for the largest multicellular land plants (which have $\sim 1000\times$ larger genomes but approximately the same numbers of chromosomes) (Figure 4.5). Over this entire gradient, a smooth, overlapping decline in recombination intensity across unicellular species, invertebrates, vertebrates, and land plants, reflects the general increase in genome sizes among these eukaryotic domains (Lynch 2007).

These observations suggest that the vast majority of the variance in the average recombination rate among eukaryotic species is simply due to variation in genome size and chromosome number. It should be noted, however, that even in the highest density genetic maps, adjacent markers are generally separated by tens of thousands to millions of base pairs, and measures of *average* levels of recombination at the genomic level need not closely reflect the features of individual chromosomal regions. Indeed, up to 100-fold differences in recombination rates can exist among regions within chromosomes, with highly localized **recombinational hotspots** existing in well-studied species (Petes 2001; de Massy 2003; Jeffreys et al. 2004; Myers et al. 2005; Arnheim et al. 2007; Coop et al. 2008; Mancera et al. 2008).

-Insert Figure 4.5 Here-

Table 4.3. Basic features of the physical and genetic maps of various eukaryotic groups, derived from a large survey of mapping studies involving high-density molecular markers. The grouping "Other unicellular species" includes algae, apicomplexans, ciliates, kinetoplastids, and oomycetes. Numbers in parentheses denote standard errors, and n denotes the number of species surveyed. Map lengths and mean chromosome sizes are in units of Morgans.

Group	Total Map Length	Genome Size (Mb)	Haploid Chr. No.	Mean Chr. Size	n
Fungi	18.3 (2.2)	36.4 (3.2)	11.9 (1.2)	1.86 (0.36)	19
Other unicellular sps.	10.9 (1.2)	80.9 (23.3)	12.9 (1.2)	0.96 (0.18)	11
Arthropods	18.1 (3.7)	679.6 (172.4)	16.1 (3.4)	1.20 (0.18)	15
Mollusks	9.2 (1.1)	1270.7 (177.2)	13.3 (1.6)	0.71 (0.09)	6
Nematodes	4.5 (1.2)	97.6 (2.5)	7.3 (1.3)	0.59 (0.05)	3
Fish	16.0 (2.3)	1185.4 (190.5)	25.1 (0.6)	0.63 (0.08)	15

Birds	23.1 (5.4)	1334.0 (48.6)	39.6 (0.4)	0.58 (0.14)	5
Mammals	23.9 (2.5)	3222.0 (108.1)	22.1 (2.2)	1.10 (0.07)	19
Angiosperms	15.9 (1.6)	2020.3 (434.2)	13.2 (0.9)	1.19 (0.07)	44

Evolution of the Recombination Rate

As in the case of the mutation rate, considerable effort has been devoted to understanding how selection might favor recombination modifiers in various contexts (e.g., Feldman et al. 1996; Barton and Otto 2005; Keightley and Otto 2006; Barton 2010; Hartfield et al. 2010). As just noted, however, the fact remains that almost all of the interspecific variation in the genome-wide amount of recombination per physical distance can be explained by a simple and largely invariant physical model of meiosis, leaving very little residual variation to be potentially assigned to mechanisms of adaptive fine-tuning. With a near universal rule of approximately one crossover per chromosome arm, one could argue that if selection is involved at all, it generally operates in a way to minimize the amount of meiotic recombination across the genome. Dumont and Payseur (2007) find that variation in recombination rates across mammalian species evolves in a manner that cannot even be discriminated from the expectations of a neutral model.

Because it minimally involves three-locus dynamics in finite populations, most population-genetic theory on the evolution of recombination-rate modifiers is highly technical and, with no simple analytical solutions available, relies heavily on computer simulations. The basic motivation underlying all such work is the general principle that natural selection often encourages the build-up of repulsion disequilibria between alleles affecting fitness, i.e., the joint accumulation of gametes with different constitutions but essentially equivalent total fitness (Chapters 5, 16). The recombinational release of such hidden genetic variance can lead to more efficient selection for joint combinations with high fitness (Chapter 7). Two general aspects of genetic systems can encourage such behavior.

First, **synergistic epistasis** (with fitness declining at an increasing rate with increasing numbers of deleterious alleles) tends to encourage the maintenance of intermediate phenotypes, thereby providing a selective advantage for recombinational production of the double mutants and their more efficient promotion/elimination by selection (Eshel and Feldman 1970; Kondrashov 1988; Charlesworth 1990; Barton 1995). On the other hand, **diminishing-returns epistasis** has the opposite effect, and encourages reduced recombination rates. As the evidence on the general incidence of these two forms of epistasis is mixed at best (Chapter 7), and the selective effects of synergistic epistasis are greatly diminished when the single-locus effects of mutations are unequal (an issue ignored in most theory, but certainly the case in reality; Butcher 1995), the role of epistasis in the evolution of recombination rates remains unclear from an empirical perspective.

Second, as already noted in Chapter 3 and further elaborated on in Chapter 7, even in the absence of nonadditive gene action, linkage reduces the efficiency of selection on multilocus systems, although the effect is expected be more pronounced in larger populations, as these will generally harbor larger numbers of cosegregating loci. Plausible arguments have been made that the power of this general feature

for the selection of modifiers that increase the recombination rate may substantially outweigh that resulting from epistasis, even when synergistic effects are common (Felsenstein and Yokoyama 1976; Otto and Barton 2001; Pálsson 2002; Barton and Otto 2005; Keightley and Otto 2006; Roze and Barton 2006).

What remains unclear is the extent to which modifiers of the recombination rate ever arise with substantial enough effects to be promoted by these kinds of associative effects. Most attempts to study the matter theoretically have focused on rather extreme situations in which either selection coefficients are very large or the magnitude of the modifier's effect on the recombination rate is extreme, and some approximations suggest that even under these conditions the selective advantage of the modifier can be quite small (Barton and Otto 2005), perhaps too small to overcome the likelihood of being lost by drift in most cases. Nevertheless, some empirical observations suggest that strong directional selection in artificial selection programs can lead to the evolution of higher recombination rates (Barton and Otto 2005), and these models may be relevant to the more general issue of the adaptive significance of sexual versus asexual reproduction, where the former entails segregation of unlinked loci as well as recombination among linked loci.

GENERAL IMPLICATIONS

The results summarized above motivate several generalizations about the intensities of mutation, recombination, and random genetic drift, in both the relative and absolute sense, and their variation across phylogenetic lineages. As these three features define the population-genetic environment within which the processes of selection occur, such knowledge provides a powerful resource for understanding the limits to molecular, genomic, and phenotypic evolution.

First, as noted above, the direct estimation of N_e in large populations is essentially impossible with current techniques. On the other hand, from information on within-population variation at putatively neutral sites, there are a number of ways to estimate the composite parameter $\theta = 4N_eu$ (or $2N_eu$ for haploids), which is equivalent to the ratio of the magnitudes of mutation and drift. As direct estimates of the mutation rate u are now available for a number of taxa, it is possible to estimate the long-term effective population size of a species by factoring the latter out from estimates of θ. For example, noting that the average estimate of θ for unicellular eukaryotes is 0.057, and that the average estimate of u for base-substitutional mutations in such species is $\sim 10^{-9}$, the average N_e for such species appears to be on the order of 3×10^7 individuals if haploidy is assumed (and half that if diploidy is assumed). These estimates are likely to be somewhat downwardly biased as selection can reduce variation at silent sites in large microbial populations. Using an average θ of 0.026 and u of 5×10^{-9} for invertebrates, average N_e for this grouping is $\sim 10^6$. Likewise, using $\theta = 0.0011$ (Example 4.1) and $u = 15 \times 10^{-9}$ (Example 4.4), long-term N_e for the human population is $\sim 19,000$.

Similar indirect inferences can be made from estimates of $\rho = 4N_ec$. For example, from Table 4.1, the average estimate of ρ for *Drosophila* species is 0.0807, whereas that for humans is ~ 0.0006, and for annual plants and long-lived trees is 0.0134 and

0.0050, respectively. From the genetic map data contributing to Figure 4.4, average c ($\times 10^{-8}$ per site per generation, based on crossovers alone) is 2.14 for *Drosophila,* 1.28 for humans, 1.59 for annual plants, and 2.93 for trees. These results imply average values of N_e of $\sim 10^6$ for *Drosophila,* 12,000 for humans, 210,000 for annual plants, and 43,000 for trees. The consistency of the results when both approaches are applied to *Drosophila* and humans is compelling.

These estimates of N_e should be considered simply as broad indicators, as θ and ρ (and therefore N_e) can vary by an order of magnitude among species within major phylogenetic groups and probably within species as well, owing to long-term temporal fluctuations (Lynch 2006). Moreover, because the mean coalescence time for a random pair of alleles is $2N_e$ generations in a diploid species (Chapter 2), polymorphism-based estimates of N_e are expected to be reasonable approximations of the average conditions experienced over only the past $\sim 2N_e$ generations. Nevertheless, several general conclusions can be made: 1) the magnitude of random genetic drift increases by a factor of nearly 10^4 from unicellular eukaryotes to large multicellular species; 2) long-term effective population sizes are generally orders of magnitudes smaller than the actual numbers of breeding adults within species, probably as a consequence of the effects of selection on mutations physically linked on chromosomes (Chapters 3 and 8); and 3) it is quite possible that no eukaryotic species, even the most enormous microbial populations, has ever had a long-term N_e much beyond 10^9, owing to the stochastic effects of selective sweeps and background selection.

Second, a long-standing puzzle in evolutionary genetics has been that the level of variation at putatively neutral sites within species is nearly independent of actual population sizes (Lewontin 1974). Given that such variation is expected to scale with N_e, and the fact that the numbers of individuals in bacterial species are many orders of magnitude greater than those for species of vertebrates and land plants, Lewontin dubbed this observation the **paradox of variation**. We now know that a strict linear increase in θ with absolute population size is unexpected owing to the effects of selection acting on linked loci. Nevertheless, given the estimates of N_e just presented, one might still expect an increase of θ on the order of 10^4 over this gradient of organisms. Yet, the observed range is only two orders of magnitude (Nei 1983; Lynch 2007; Leffler et al. 2012).

The reason for this discrepancy is made clear by the summary above. The mutation rate u is not independent of N_e, but instead strongly declines with increasing N_e, thereby partly compensating for the direct influence of N_e on θ. As a consequence, it appears that in no species does the power of mutation exceed that of random genetic drift (i.e., θ is always much smaller than 1.0). Moreover, because estimates of ρ are also always well below 1.0, the same conclusion can be drawn with respect to the relative magnitudes of recombination per nucleotide site and the power of random genetic drift.

Third, because the per-site mutation rate increases with genome size (Figure 4.4), whereas the per-site meiotic crossover rate declines (Figure 4.5), it can be concluded that the ratio of the power of mutation to that of recombination increases substantially with genome size. Using the regression relationships in these two figures, for eukaryotic genomes of size 10, 100, 10^3, and 10^4 Mb, respectively, average u/c is approximately 0.00076, 0.028, 1.01, and 36.5. These extrapolations are consis-

tent with the indirect (polymorphism-based) estimates of u/c implied in Table 4.1, which are subject to substantial sampling error but fall in the order-of-magnitude range of 1 to 100 for animals and land plants (with genome sizes in the range of 100 to 10^4 Mb).

These ideas need to be tempered by the fact that for closely spaced sites, gene conversion causes the recombination rate to be elevated relative to that expected on the basis of crossing over alone. From Equations (4.16b,c), the degree of inflation is $\simeq (x+2)/x$, where x is the fraction of recombination events resulting in a crossover. As x is typically in the neighborhood of 0.1 (Lynch et al. in prep.), this implies that the effective value of u/c may be as much as $10\times$ lower than the values suggested above. On the other hand, with the emerging data suggesting that most recombination events are concentrated at a small number of hotspots, the recombination rate at most nucleotide sites will be much lower than the average, implying that except near hotspots u/c will be higher than implied with the use of average c values.

Literature Cited

Anderson, E. C. 2005. An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. *Genetics* 170: 955–967. [4]

Anderson, E. C., E. G. Williamson, and E. A. Thompson. 2000. Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* 156: 2109–2118. [4]

Andolfatto, P., and M. Nordborg. 1998. The effect of gene conversion on intralocus associations. *Genetics* 148: 1397–1399. [4]

Andolfatto, P., and M. Przeworski. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila. Genetics* 156: 257–268. [4]

André, J. B., and B. Godelle. 2006. The evolution of mutation rate in finite asexual populations. *Genetics* 172: 611–626. [4]

Angerer, W. P. 2001a. A note on the evaluation of fluctuation experiments. *Mutat. Res.* 479: 207–224. [4]

Angerer, W. P. 2001b. An explicit representation of the Luria-Delbrück distribution. *J. Math. Biol.* 42: 145–174. [4]

Arnheim, N., P. Calabrese, and I. Tiemann-Boege. 2007. Mammalian meiotic recombination hot spots. *Annu. Rev. Genet.* 41: 369–399. [4]

Arunyawat, U., W. Stephan, and T. Städler. 2007. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol. Biol. Evol.* 24: 2310–2322. [4]

Baer, C. F., M. M. Miyamoto, and D. R. Denver. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8: 619–631. [4]

Barton, N. H. 1995. A general model for the evolution of recombination. *Genet. Res.* 65: 123–145. [4]

Barton, N. H. 2010. Mutation and the evolution of recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365: 1281–1294. [4]

Barton, N. H., and S. P. Otto. 2005. Evolution of recombination due to random drift. *Genetics* 169: 2353–2370. [4]

Beaumont, M. A. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* 164: 1139–1160. [4]

Berthier, P., M. A. Beaumont, J. M. Cornuet, and G. Luikart. 2002. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* 160: 741–751. [4]

Bollback, J. P., T. L. York, and, R. Nielsen. 2008. Estimation of $2N_e s$ from temporal allele frequency data. *Genetics* 179: 497–502. [4]

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc. Natl. Acad. Sci. USA* 101: 15255–15260. [4]

Butcher, D. 1995. Muller's ratchet, epistasis and mutation effects. *Genetics* 141: 431–437. [4]

Campbell, C. D., et al. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44: 1277-1281. [4]

Charlesworth, B. 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.* 55: 199–221. [4]

Chen, H., P. L. Morrell, M. de la Cruz, and M. T. Clegg. 2008. Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). *J. Hered.* 99: 382–389. [4]

Clune, J., D. Misevic, C. Ofria, R. E. Lenski, S. F. Elena, and R. Sanjuán. 2008. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. *PLoS Comput. Biol.* 4: e1000187. [4]

Conrad, D. F., et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43: 712–714. [4]

Coop, G., X. Wen, C. Ober, J. K. Pritchard, and M. Przeworski. 2008. High-resolution mapping of crossovers reveals extensive variation in fine–scale recombination patterns among humans. *Science* 319: 1395–1398. [4]

Crow, J. F. 2000. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* 1: 40–47. [4]

Dawson, K. J. 1999. The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theor. Pop. Biol.* 55: 1–22. [4]

de Massy, B. 2003. Distribution of meiotic recombination sites. *Trends Genet.* 19: 514–522. [4]

Denamur, E., and I. Matic. 2006. Evolution of mutation rates in bacteria. *Mol. Microbiol.* 60: 820–827. [4]

Denver, D. R., K. Morris, M. Lynch, and W. K. Thomas. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430: 679–682. [4]

Denver, D. R., L. J. Wilhelm, D. K. Howe, K. Gafner, P. C. Dolan, and C. F. Baer. 2012. Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. *Genome Biol. Evol.* 4: 513–522. [4]

Desai, M. M., and D. S. Fisher. 2011. The balance between mutators and nonmutators in asexual populations. Genetics 188: 997-1014. [4]

Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA* 88: 7160–7164. [4]

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. *Genetics* 148: 1667–1686. [4]

Drost, J. B., and W. R. Lee. 1995. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila,* mouse, and human. *Environ. Mol. Mutagen.* 25 Suppl 26: 48–64. [4]

Dumont, B. L., and B. A. Payseur. 2008. Evolution of the genomic rate of recombination in mammals. *Evolution* 62: 276–294. [4]

Eshel, I., and M. W. Feldman. 1970. On the evolutionary effect of recombination. *Theor. Popul. Biol.* 1: 88–100. [4]

Ethier, S. N., and R. C. Griffiths. 1990. On the two-locus sampling distribution. *J. Math. Biol.* 29: 131–159. [4]

Fearnhead, P., and P. Donnelly. 2001. Estimating recombination rates from population genetic data. *Genetics* 159: 1299–1318. [4]

Feldman, M. W., S. P. Otto, and F. B. Christiansen. 1996. Population genetic perspectives on the evolution of recombination. *Annu. Rev. Genet.* 30: 261–295. [4]

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68: 581–597. [4]

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.* 59: 139–147. [4]

Felsenstein, J., and S. Yokoyama. 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. *Genetics* 83: 845–859. [4]

Flury, C., M. Tapio, T. Sonstegard, C. Drögemüller, T. Leeb, H. Simianer, O. Hanotte, and S. Rieder. 2010. Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *J. Anim. Breed. Genetics* 127: 339–337. [4]

Frankham, R. 1995. Effective population size / adult population size ratios in wildlife: a review. *Genet. Res.* 66: 95–107. [4]

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Amer. J. Hum. Genet.* 69: 831–843. [4]

Fu, Y.-X. 1994a. A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136: 685–692. [4]

Fu, Y.-X. 1994b. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* 138: 1375–1386. [4]

Fu, Y.-X. 1995. Statistical properties of segregating sites. *Theor. Pop. Biol.* 48: 172–197. [4]

Fu, Y.-X., and W.-H. Li. 1993a. Maximum likelihood estimation of population parameters. *Genetics* 134: 1261–1270. [4]

Fu, Y.-X., and W.-H. Li. 1993b. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709. [4]

Fujimoto, A., T. Kado, H. Yoshimaru, Y. Tsumura, and H. Tachida. 2008. Adaptive and slightly deleterious evolution in a conifer, *Cryptomeria japonica*. *J. Mol. Evol.* 67: 201–210. [4]

Golding, G. B. 1984. The sampling distribution of linkage disequilibrium. *Genetics* 108: 257–274. [4]

Goldringer, I., and T. Bataillon. 2004. On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics* 168: 563–568. [4]

Haag-Liautard, C., M. Dorris, X. Maside, S. Macaskill, D. L. Halligan, D. Houle, B. Charlesworth, and P. D. Keightley. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445: 82–85. [4]

Haldane, J. B. S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J. Genetics* 8: 299–309. [4]

Hamblin, M. T., M. G. Salas Fernandez, A. M. Casa, S. E. Mitchell, A. H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. *Genetics* 171: 1247–1256. [4]

Hartfield, M., S. P. Otto, and P. D. Keightley. 2010. The role of advantageous mutations in enhancing the evolution of a recombination modifier. *Genetics* 184: 1153–1164. [4]

Haubold, B., P. Pfaffelhuber, and M. Lynch. 2010. mlDiv – A program for estimating the population mutation and recombination rates from shotgun-sequenced genomes. *Mol. Ecol.* 19, Suppl. 1: 277–284. [4]

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13: 635–643. [4]

Hedrick, P. 2005. Large variance in reproductive success and the N_e/N ratio. *Evolution* 59: 1596–1599. [4]

Herr, A. J., M. Ogawa, N. A. Lawrence, L. N. Williams, J. M. Eggington, M. Singh, R. A. Smith, and B. D. Preston. 2011. Mutator suppression and escape from replication error-induced extinction in yeast. *PLoS Genet.* 7(10): e1002282. [4]

Hill, W. G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor. Pop. Biol.* 8: 117–126. [4]

Hill, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* 38: 209–216. [4]

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* 50: 245–250. [4]

Hudson, R. R. 2001. Two-locus sampling distributions and their application. *Genetics* 159: 1805–1817. [4]

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. [4]

Jain, K., and A. Nagar. 2013. Fixation of mutators in asexual populations: the role of genetic drift and epistasis. Evolution 67: 1143–1154. [4]

Jeffreys, A. J., J. K. Holloway, L. Kauppi, C. A. May, R. Neumann, M. T. Slingsby, and A. J. Webb. 2004. Meiotic recombination hot spots and human DNA diversity. *Phil. Trans. Roy. Soc. Lond. B Biol. Sci.* 359: 141–152. [4]

Jenkins, P. A., and Y. S. Song. 2009. Closed-form two-locus sampling distributions: accuracy and universality. *Genetics* 183: 1087–1103. [4]

Johnson, P. L., and M. Slatkin. 2008. Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25: 199–206. [4]

Johnson, T. 1999a. The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. Biol. Sci.* 266: 2389–2397. [4]

Johnson, T. 1999b. Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. *Genetics* 151: 1621–1631. [4]

Kang, C. J., and P. Marjoram. 2011. Inference of population mutation rate and detection of segregating sites from next-generation sequence data. *Genetics* 189: 595–605. [4]

Keightley, P. D., and D. L. Halligan. 2011. Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* 188: 931–940. [4]

Keightley, P. D., and S. P. Otto. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443: 89–92. [4]

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana. Nature Genetics* 39: 1151–1155. [4]

Kimble, J., and S. Ward. 1998. Germ-line development and fertilization. *In* W. B. Wood (ed.), *The nematode Caenorhabditis elegans,* pp. 191–213. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. [4]

Kimura, M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* 9: 23–34. [4]

Kondrashov, A. S. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* 336: 435–440. [4]

Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* 21: 12–27. [4]

Kong, A., et al. 2012. Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* 488: 471-475. [4]

Krimbas, C. B., and S. Tsakas. 1971. The genetics of *Dacus oleae.* V. Changes of esterase polymorphism in a natural population following insecticide control-selection or drift? *Evolution* 25: 454–460. [4]

Kuhner, M. K. 2009. Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* 24: 86–93. [4]

Kuhner, M. K., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* 156: 1393–1401. [4]

Lagercrantz, U., M. K. Osterberg, and M. Lascoux. 2002. Sequence variation and haplotype structure at the putative flowering-time locus COL1 of *Brassica nigra. Mol. Biol. Evol.* 19: 1474–1482. [4]

Lang, G. I., and A. W. Murray. 2008. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae. Genetics* 178: 67–82. [4]

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman. 2000. Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome. *Genetics* 156: 1837–1852. [4]

Lea, D. E., and C. A. Coulson. 1949. The distribution of the number of mutants in bacterial populations. *J. Genetics* 28: 264–285. [4]

Lefebvre, J. F., and D. Labuda. 2008. Fraction of informative recombinations: a heuristic approach to analyze recombination rates. *Genetics* 178: 2069–2079. [4]

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, and M. Przeworski. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* 10(9):e1001388.

Leigh, E. G., Jr. 1970. Natural selection and mutability. *Amer. Natur.* 104: 301–305. [4]

Leigh, E. G., Jr. 1973. The evolution of mutation rates. *Genetics* (Suppl.) 73: 1–18. [4]

Lewontin, R. C. 1974. *The genetic basis of evolutionary change.* Columbia Univ. Press, New York, NY. [4]

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175–195. [4]

Li, W.-H., and Y.-X. Fu. 1999. Coalescent theory and its application in population genetics. *In* M. E. Halloran and S. Geisser (eds.), *Statistics in genetics,* pp. 45–79. Springer Verlag, New York. [4]

Liu, A., and J. M. Burke. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* 173: 321–330. [4]

Liu, X., and Y.-X. Fu. 2008. Algorithms to estimate the lower bounds of recombination with or without recurrent mutations. *BMC Genomics* 9 (Suppl. 1): S24. [4]

Luikart, G., and J. M. Cornuet. 1999. Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* 151: 1211–1216. [4]

Luria, S. E., and M. Delbrück. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28: 491–511. [4]

Lynch, M. 2006. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 23: 450–468. [4]

Lynch, M. 2007. *The origins of genome complexity.* Sinauer Assocs., Inc. Sunderland, MA. [4]

Lynch, M. 2008a. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* 25: 2409–2419. [4]

Lynch, M. 2008b. The cellular, developmental, and population-genetic determinants of mutation-rate evolution. *Genetics* 180: 933–943. [4]

Lynch, M. 2009a. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 182: 295–301. [4]

Lynch, M. 2009b. Rate, molecular spectrum, and consequences of spontaneous mutations in man. *Proc. Natl. Acad. Sci. USA* 107: 961–968. [4]

Lynch, M. 2010. Evolution of the mutation rate. *Trends Genetics* 26: 345–352. [4]

Lynch, M. 2011. The lower bound to the evolution of mutation rates. *Genome Biol. Evol.* 3: 1107–1118. [4]

Lynch, M., L. M. Bobay, F. Catania, J.-F. Gout, and M. Rho. 2011. The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet.* 12: 347–366. [4]

Lynch, M., and T. J. Crease. 1990. The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* 7: 377–394. [4]

Lynch, M., W. Sung, K. Morris, N. Crown, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA* 105: 9272–9277. [4]

Lynch, M., S. Xu, T. Maruki, X. Jiang, P. Pfaffelhuber, and B. Haubold. Genome-wide linkage-disequilibrium profiles from single individuals. (in prep.) [4]

Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454: 479–485. [4]

Massey, S. E. 2008. The proteomic constraint and its role in molecular evolution. *Mol. Biol. Evol.* 25: 2557-2565. [4]

Mather, K. A., A. L. Caicedo, N. R. Polato, K. M. Olsen, S. McCouch, and M. D. Purugganan. 2007. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177: 2223–2232. [4]

McEvoy, B. P., J. E. Powell, M. E. Goddard, and P. M. Visscher. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.* 21: 821–829. [4]

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160: 1231–1241. [4]

Messer, P. W. 2009. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* 182: 1219–1232. [4]

Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. *Genetics* 173: 1705–1723. [4]

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310: 321–324. [4]

Myers, S. R., and R. C. Griffiths. 2003. Bounds on the minimum number of recombination events in a sample history. *Genetics* 163: 375–394. [4]

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156: 297–304. [4]

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583–590. [4]

Nei, M. 1983. Genetic polymorphism and the role of mutation in evolution. *In* M. Nei and

R. K. Koehn (eds.), *Evolution of genes and proteins,* pp. 165–190. Sinauer Assocs., Inc., Sunderland, MA. [4]

Nei, M., and L. Jin. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* 6: 290–300. [4]

Nei, M., and A. K. Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76: 379–390. [4]

Nei, M., and F. Tajima. 1981a. DNA polymorphism detectable by restriction endonucleases. *Genetics* 97: 145–163. [4]

Nei, M., and F. Tajima. 1981b. Genetic drift and estimation of effective population size. *Genetics* 98: 625–640. [4]

Nei, M., and F. Tajima. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. *Genetics* 105: 207–217. [4]

Ness, R. W., A. D. Morgan, N. Colegrave, and P. D. Keightley. 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii. Genetics* 192: 1447-1454. [4]

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154: 931–942. [4]

Nishant, K. T., et al. 2010. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genet.* 6(9). pii: e1001109. [4]

Notley-McRobb, L., S. Seeto, and T. Ferenci. 2002. Enrichment and elimination of mutY mutators in *Escherichia coli* populations. *Genetics* 162: 1055–1062. [4]

Ohta, T., and M. Kimura. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63: 229–238. [4]

Otto, S. P., and N. H. Barton. 2001. Selection for recombination in small populations. *Evolution* 55: 1921–1931. [4]

Ossowski, S., K. Schneeberger, J. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, and M. Lynch. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana. Science* 327: 92–94. [4]

Pálsson, S. 2002. Selection on a modifier of recombination rate due to linked deleterious mutations. *J. Hered.* 93: 22–26. [4]

Palstra, F. P., and D. E. Ruzzante. 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Mol. Ecol.* 17: 3428–3447. [4]

Pamilo, P., and S. L. Varvio-Aho. 1980. On the estimation of population size from allele frequency changes. *Genetics* 95: 1055–1057. [4]

Petes, T. D. 2001. Meiotic recombination hot spots and cold spots. *Nat. Rev. Genetics* 2: 360–369. [4]

Pluzhnikov, A., and P. Donnelly. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144: 1247–1262. [4]

Pollak, E. 1983. A new method for estimating the effective population size from allele frequency changes. *Genetics* 104: 531–548. [4]

Ptak, S. E., K. Voelpel, and M. Przeworski. 2004. Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics* 167: 387–397. [4]

Pudovkin, A. I., D. V. Zaykin, and D. Hedgecock. 1996. On the potential for estimating the effective number of breeders from heterozygote–excess in progeny. *Genetics* 144: 383–387. [4]

Pyhäjärvi, T., M. R. Garca–Gil, T. Knürr, M. Mikkonen, W. Wachowiak, and O. Savolainen. 2007. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* 177: 1713–1724. [4]

Robertson, A. 1965. The interpretation of genotypic ratios in domestic animal populations. *Anim. Prod.* 7: 319–324. [4]

Rosche, W. A., and P. L. Foster. 2000. Determining mutation rates in bacterial populations. *Methods* 20: 4–17. [4]

Roze, D., and N. H. Barton. 2006. The Hill-Robertson effect and the evolution of recombination. *Genetics* 173: 1793–1811. [4]

Sarkar, S., W. T. Ma, and G. H. Sandri. 1992. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* 85: 173–179. [4]

Schrider, D. R., D. Houle, M. Lynch, and M. W. Hahn. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster. Genetics* (in press). [4]

Seyfert, A. L., M. E.A. Cristescu, L. Frisse, S. Schaack, W. K. Thomas, and M. Lynch. 2008. The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex. Genetics* 178: 2113–2121. [4]

Sniegowski, P. D., P. J. Gerrish, T. Johnson, and A. Shaver. 2000. The evolution of mutation rates:separating causes from consequences. *Bioessays* 22: 1057–1066. [4]

Stephens, J. C. 1986. On the frequency of undetectable recombination events. *Genetics* 112: 923–926. [4]

Sturtevant, A. H. 1937. Essays on evolution. I. On the effects of selection on mutation rate. *Quart. Rev. Biol.* 12: 464–476. [4]

Sung, W., M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. USA* 109: 18488–18492. [4]

Sved, J. A., A. F. McRae, and P. M. Visscher. 2008. Divergence between human populations estimated from linkage disequilibrium. *Amer. J. Hum. Gen.* 83: 737–743. [4]

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460. [4]

Tajima, F., and M. Nei. 1984. Note on genetic drift and estimation of effective population size. *Genetics* 106: 569–574. [4]

Tallmon, D. A., G. Luikart, and M. A. Beaumont. 2004. Comparative evaluation of a new effective population size estimator based on approximate bayesian computation. *Genetics* 167: 977–988. [4]

Tenaillon, M. I., J. U'Ren, O. Tenaillon, and B. S. Gaut. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* 21: 1214–1225. [4]

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17: 520–526. [4]

Tröbner, W., and R. Piechocki. 1984 Selection against hypermutability in *Escherichia coli* during long term evolution.*Mol. Gen. Genet.* 198: 177–178. [4]

Wakeley, J. 1997. Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* 69: 45–48. [4]

Wall, J. D. 2000. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* 17: 156–163. [4]

Wang, J. 2001. A pseudo–likelihood method for estimating effective population size from temporally spaced samples. *Genet. Res.* 78: 243–257. [4]

Wang, J. 2005. Estimation of effective population sizes from data on genetic markers. *Phil. Trans. Roy. Soc. Lond. B Biol. Sci.* 360: 1395–1409. [4]

Wang, J. 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Mol. Ecol.* 18: 2148–2164. [4]

Waples, R. S. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121: 379–391. [4]

Waples, R. S., and M. Yokota. 2007. Temporal estimates of effective population size in species with overlapping generations. *Genetics* 175: 219–233. [4]

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7: 256–276. [4]

Wielgoss, S., J. E. Barrick, O. Tenaillon, M. J. Wiser, W. J. Dittmar, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci USA 110: 222–227. [4]

Weir, B. S., and W. G. Hill. 1980. Effect of mating structure on variation in linkage disequilibrium. *Genetics* 95: 477–488. [4]

Wilke, C. O., J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412: 331–333. [4]

Williams, L. N., A. J. Herr, and B. D. Preston. 2013. Emergence of DNA polymerase ϵ antimutators that escape error-induced extinction in yeast. *Genetics* 193: 751-770. [4]

Williamson, E. G., and M. Slatkin. 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* 152: 755–761. [4]

Xue, Y., et al. 2009. Human Y chromosome base–substitution mutation rate measured by direct sequencing in a deep–rooting pedigree. *Curr. Biol.* 19: 14531457. [4]

Yu, N., M. I. Jensen–Seaman, L. Chemnick, J. R. Kidd, A. S. Deinard, O. Ryder, K. K. Kidd, and W.-H. Li. 2003. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* 164: 1511–1518. [4]

Zietkiewicz, E., et al. 2003. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *Amer. J. Hum. Genet.* 73: 994–1015. [4]

Figure 4.1. Expected sampling standard deviations for estimates of θ from sequences assumed to be neutral, in drift-mutation equilibrium, and experiencing no intragenic recombination. Results are derived from Equations 4.2 (Tajima estimator based on heterozygosity), 4.4 (Watterson estimator based on segregating sites), and 4.5 (maximally efficient), for θ =0.1, 0.01, and 0.001 in descending order. The assumed number of sites is $L = 10,000$ in all cases.

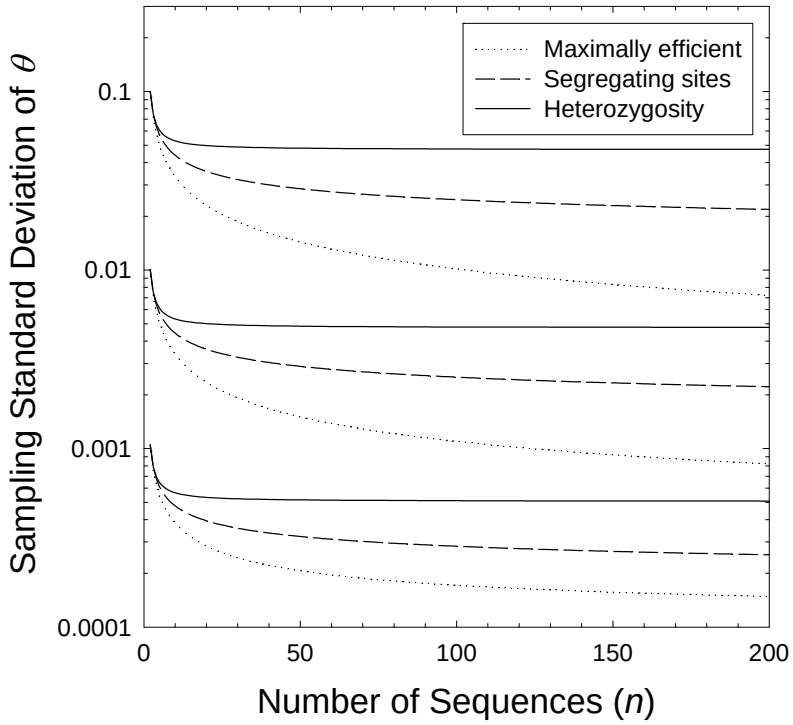

Figure 4.2. The upper and lower bounds on the fraction of recombination events that produce nonparental gametes among two or more neutrally evolving sites (from Equations 4.8a,b). $\Theta = L\theta$ is the product of the population mutation rate per site (θ) and the length of the segment (L, in base pairs).

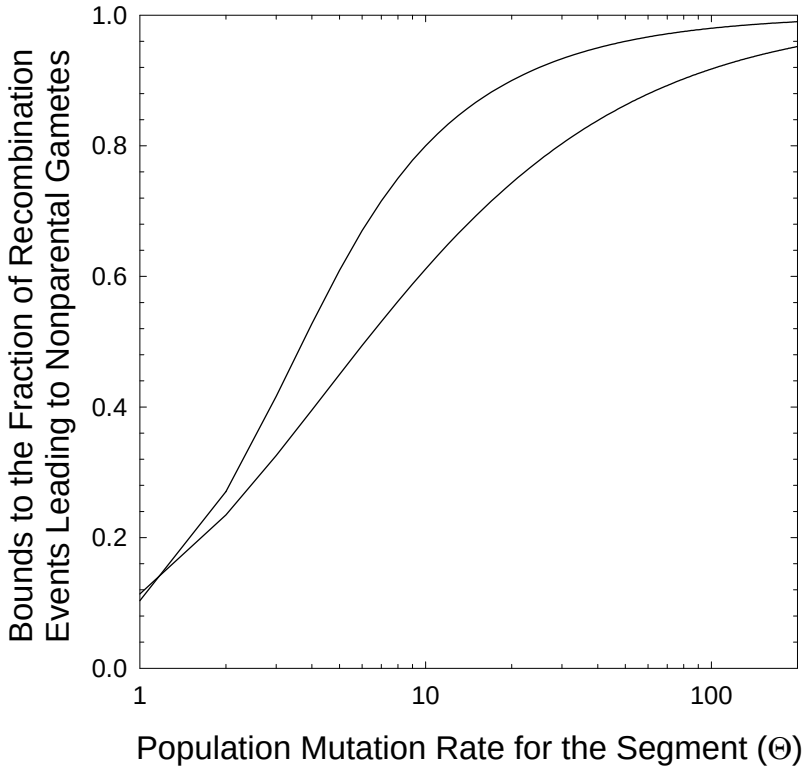

Figure 4.3. Estimates of historical values of N_e using linkage-disequilibrium between large numbers of pairs of markers with different genetic-map distances. The estimates were pooled into categories with different values of c (between markers), with the bin-specific values of $1/(2c)$ serving as estimates of the age (in generations) for which the category provides an estimate of N_e, with the latter being calculated by using the simplified version of Equation 4.14 given the average estimate of r_D^2 and c for the bin. a) Estimates of historical changes in N_e for a Utah population of European extraction. For a given generational time slice, each point represents an estimate based on the markers from each of the 22 human autosomes. Note the rapid increase in N_e in the recent past. After Tenesa et al. (2007). b) Estimates for the Swiss Eringer breed of cattle. Here, the different curves represent different assumptions used to correct estimates of ρ for sampling effects and different estimates of the fine-scale recombination rates. Regardless of the assumptions involved, it is clear that in contrast to the results for the human population, the N_e has dramatically declined over the past 500 years. After Flury et al. (2010).

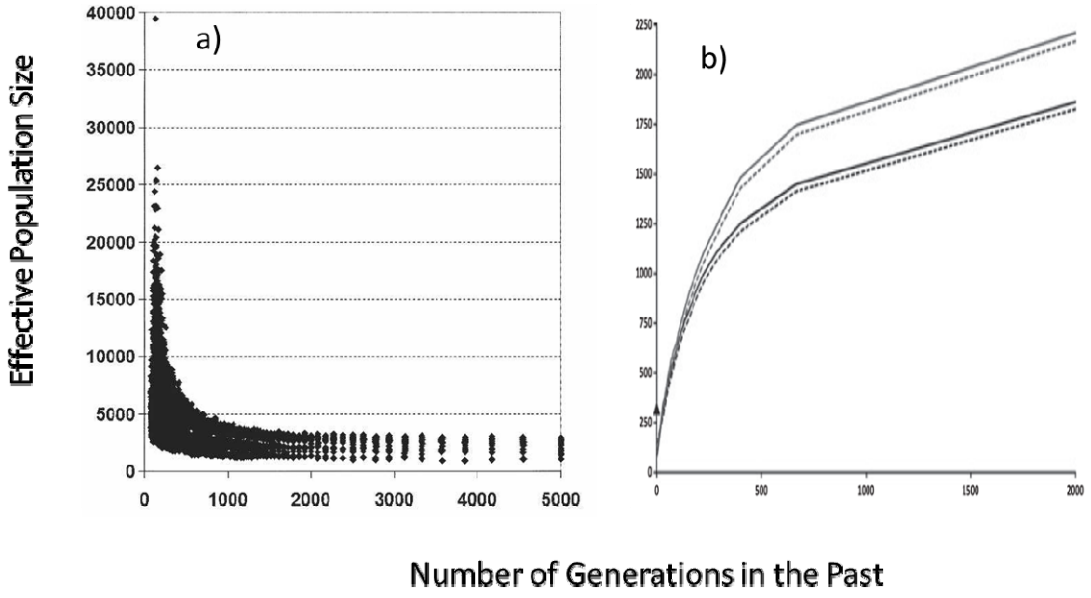

Figure 4.4. Scaling of the mutation rate (u, for base-substitutional changes only, in units of 10^{-9} per nucleotide site per generation) as a function of genome size (G, in units of 10^6 nucleotide sites). The least-squares regression describing the overall relationship is $u = 0.16G^{0.66}$. Data are all from Lynch (2010) and Sung et al. (2012).

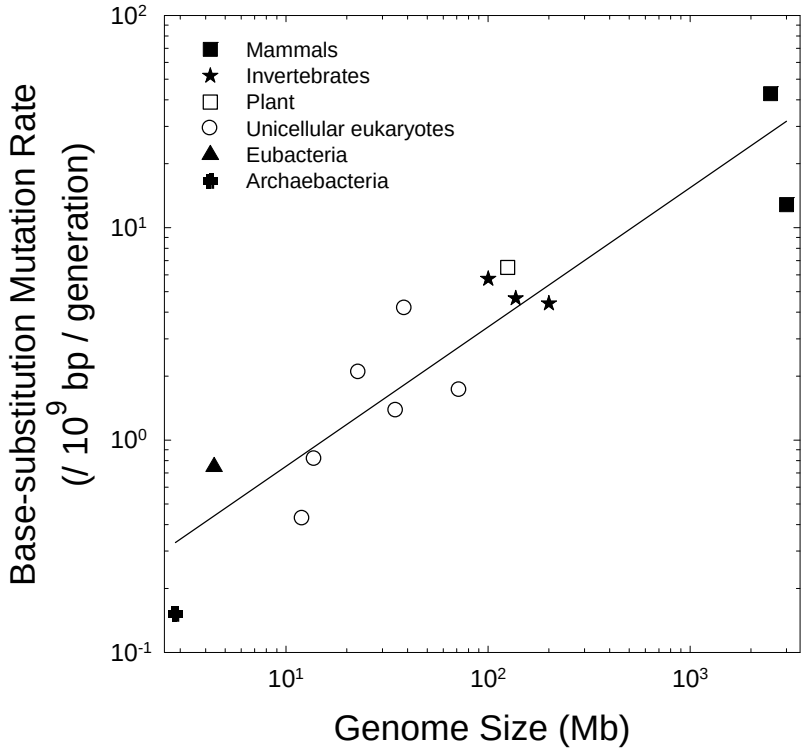

Figure 4.5. Average rates of recombination per physical distance for four major groupings of eukaryotes, determined from information on total physical and genetic map sizes. The two dashed lines have slopes of -1.0 in accordance with the theory discussed in the text. Letting x be the average number of crossovers, and N be the number of chromosomes, the top line assumes $xN = 50$, i.e., 50 chromosomes with an average length of 1.0 Morgans, 25 with average lengths of 2.0 Morgans, 100 with average lengths of 0.5 Morgans, etc. The lower line assumes $xN = 3$. For the plotted species, x is in the range of 0.3 to 3.1 (with one exception) and N is in the range of 3 to 44. From Lynch et al. (2011).