

4

THE NONADAPTIVE FORCES OF EVOLUTION

Draft Version 24 August 2008

Although natural selection plays a major role in the evolution of many quantitative traits, three additional factors determine the patterns of genetic variation within and among populations upon which natural selection operates. 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 might be added to this list, although we regard this added complexity as being independent of the internal genetic machinery in a population, and defer attention to this matter until later chapters that consider the consequences of population subdivision.

In the absence of external forces of 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. This logic works in reverse as well – 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, two-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 information resulting from such analyses, providing information that will play a central role in applications of the theory derived 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 numerous refinements in the near future.

Although our ultimate desire is to procure accurate estimates of the forces of mutation, recombination, and drift in a wide array of species, it is often much easier to obtain ratios of two of these features than to measure the individual parameters. However, this is not necessarily 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 and the power of drift defines the distribution of linkage disequilibrium. Therefore, we will first consider methods for estimating $\theta = 4N_e u$ and $\rho = 4N_e c$, before summarizing the approaches for estimating N_e , u , and c separately. It will be shown that although methods exist for the direct measurement of N_e , accurate estimates are difficult to achieve, especially in large populations. However, by using the combined estimates of θ , ρ , u , and c , reliable measures of long-term N_e can be obtained indirectly.

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 Chapter 8. In addition, we will largely focus on measures at the level of individual nucleotide sites, as it is now straight-forward 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.

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 locus (or nucleotide site) 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 for the relevant unit of analysis (in this case, nucleotide sites). As will be seen below, average heterozygosity per neutral nucleotide site is generally far below 1.0 in all phylogenetic groups, in which case 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 is $2N_e u$, yielding the same interpretation).

Nucleotide Diversity

Suppose n random sequences have been obtained from multicellular individuals for a particular genomic region. In principle, such a region might consist of intronic or intergenic DNA or of the subset of silent (synonymous) sites in a coding region. 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, then

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{\substack{i=1 \\ j>i}}^n k_{ij} / L \quad (4.1)$$

provides a heterozygosity-based estimate of $4N_e u$ (Tajima 1983). If we are to make such an interpretation confidently, aside from the issue of whether the assumptions

of neutrality and equilibrium are valid, it is critical to know the sampling variance of $\hat{\theta}_H$, which results from two sources.

First, *evolutionary variance* exists for θ_H in that nucleotide frequencies fluctuate over time as a consequence of the stochastic forces of mutation and drift. This matter has been addressed previously in Chapter 2. Assuming a population in drift-mutation equilibrium, the expected evolutionary variance of the average value of θ_H over L independent (effectively unlinked) sites is $\sim (\theta/3)[(2\theta/3) + (1/L)]$ (Tajima 1983). This source of variance depends only on the internal features of the population, and is independent of N_e . Second, *sampling variance* results from the reliance of estimates on a finite number of sampled sequences, which for an equilibrium population is $\sim 2\theta/[3(n-1)][(2n+3)\theta/(3n)] + (1/L)$, where n is the number of alleles 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

$$\sigma^2(\hat{\theta}_H) \simeq \frac{\theta}{3(n-1)} \left(\frac{2(n^2 + n + 3)\theta}{3n} + \frac{n+1}{L} \right) \quad (4.2)$$

(Pluzhnikov and Donnelly 1996). This relationship shows that the variance of estimates of θ based on nucleotide diversity approaches a minimum of $[\theta/(3L)] + (2\theta^2/9)$ with increasing numbers of sampled alleles per site, i.e., as $n \rightarrow \infty$. Even with an enormous amount of sequence per individual (large L), the sampling coefficient of variation of $\hat{\theta}_H$ is no lower than $\sqrt{2/9}$ due to the fact that the genealogical structure of populations results in the nonindependence of samples drawn at any particular point of time.

Here it is worth reemphasizing that Equation 4.2 is only an appropriate estimator of the variance of a nucleotide-diversity estimate if the latter is based on neutral sites in drift-mutation equilibrium. Even for assuredly neutral sites, this expression will not apply for nonequilibrium situations, e.g., populations that have experienced relatively recent expansions or contractions. For such situations, the variance of heterozygosity must be evaluated more directly from the spectrum of allele frequencies across all sites and higher-order moments of them in order to account for the variance of allele sampling within sites and the variance of heterozygosity among sites. A number of 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 hundreds of thousands to 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 can be measured, providing a direct estimate of the evolutionary sampling 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 2008). Given the increasing economic feasibility of sequencing individual genomes, these observations are 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 sequences, putting the effort instead into sampling more sites, i.e., maximizing L at the expense of n .

Number of Segregating Sites

Although nucleotide diversity may be the most transparent means of estimating θ , it is by no means the only approach or even the most efficient one. 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 more than a single allele, S clearly increases with the length of the sequence and the number of individuals assayed, but Watterson (1975) showed that under the assumptions of neutrality and drift-mutation equilibrium, a nonbiased estimator of θ is

$$\hat{\theta}_S = S/(La_n) \quad (4.3)$$

where $a_n = \sum_{j=1}^{n-1} 1/j$.

Watterson and Tajima actually focused on per-locus measures of θ , so L does not appear in their expressions, which have been modified here to Equations 4.1 and 4.3. However, the central point is that when the nucleotide sites surveyed are neutral and in drift-mutation equilibrium, both of these expressions provide estimates of the per-site parameter $\theta = 4N_e u$. In Chapter 8, we will see that when the assumptions of neutrality and/or equilibrium are violated, the relative values of θ_H and θ_S deviate from each other in ways that yield insight into past population-genetic processes.

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

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

where $b_n = \sum_{j=1}^{n-1} 1/j^2$. 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 reemphasized that both Equations 4.2 and 4.4 were derived under the assumptions of sequences experiencing negligible recombination. The necessary modifications to allow for intragenic recombination occurs are derived in Pluzhnikov and Donnelly (1996; their Equations 6 and 7), and 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 , and such effects can be quite deceptive because rare alleles are expected to be common under the neutral hypothesis. Johnson and Slatkin (2008) have suggested a means for eliminating the bias from S when an accurate estimate of the sequencing-error rate is available.

Alternative Approaches

Felsenstein (1992) pointed out that neither of the above approaches are likely to provide the most efficient estimates of θ (i.e., estimates with minimum sampling variance), as they do not fully utilize all of the information in the sample of sequences. In particular, both approaches ignore the phylogenetic relationships of sequences, although as shown in Chapter 2, under neutrality the expected genealogical branch lengths can be expressed in terms of θ . To evaluate how much improvement might be achieved by exploiting such information, Fu and Li (1993) 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. In this case, the expected sampling variance of the estimator is

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

Comparison of the behavior of the optimal estimator with Equations 4.2 and 4.4 demonstrates 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 mutation sizes (number of sequences in the sample that contain derived copies of the mutation) and types (number of sites with i or $n-i$ copies of a variant) (Fu 1995; Li and Fu 1999), asymptotically perform in a near optimal manner as the sample size increases.

Empirical Observations

Estimates of θ , mostly derived as silent-site heterozygosity from protein-coding genes as defined by Equation 4.1, have been summarized for a wide range of species by Lynch (2007). Across a diverse assemblage of > 100 eukaryotic and prokaryotic species, there is an inverse relationship between organism size and θ_H , 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 value for prokaryotes is nearly twice that for unicellular eukaryotes (mean = 0.057, SD = 0.078), and estimates for invertebrates (mean = 0.026, SD = 0.015), 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 below), but because of sampling error at the gene, individual, and population levels, the standard deviations must be upwardly biased with respect to true species-specific variation.

As estimators of $4N_e u$ ($2N_e u$ for haploids), silent-site heterozygosity measures for both of the unicellular groups are likely to be downwardly biased 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, the variation at silent sites will under-represent the neutral expectation if such sites experience some form of purifying selection. Such conditions can arise as a consequence of translation-associated selection when certain tRNAs have higher affinities for certain alternative codons (through differential tRNA abundance and/or physical features), or when double-strand break repair is inhibited between highly divergent alleles. The molecular biological underpinnings of both of these factors, as well as their potential population-genetic consequences, are reviewed in Lynch (2007). Because both forms of selection are quite weak, they will be most effective in populations with very large N_e , and the central conclusion is that although θ_H may underestimate $4N_e u$ ($2N_e u$ for haploids) in some microbial species by as much as tenfold, the bias is most likely minor in multicellular eukaryotes.

Taken together, 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, and in large multicellular land plants and vertebrates, the ratio is almost always on the order of 0.03 or 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 that N_e is usually substantially smaller than the actual number of reproductive individuals in a population.

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 individual 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 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. Whereas 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), patterns of linkage disequilibrium (LD) over kilobase-length regions have the potential to reveal much more refined views of the recombinational landscape averaged over longer evolutionary time scales. Thus, we will initially focus on the latter before making connections with the results derived from linkage maps.

Just as the amount of segregating variation at neutral sites provides insight into the population mutation parameter $\theta = 4N_e u$, the amount of standing LD is a function of the population recombination rate, $\rho = 4N_e c$ (Chapter 2). Although a wide variety of methods for estimating ρ have been proposed, the challenges to obtaining accurate estimates are formidable. The markers employed must not only have at least moderate frequencies (to ensure the procurement of accurate estimates of gamete frequencies), but also be neutral (to ensure the validity of the application of drift-recombination theory). Moreover, aside from the fact that most of the proposed estimators are quite technical, they are also saddled with very high sampling variance. The latter is not simply a consequence of inadequate statistical procedures, but rather of the nature of the recombinational process itself, in particular the limited detectability of historical recombination events. Although it is relatively straight-forward to identify mutations from patterns of allelic divergence, a substantial fraction of recombination events leave no trace. Such is the case for all recombination events between two adjacent markers in parents that are not doubly heterozygous. Thus, the procurement of unbiased estimates of ρ requires a method for accounting for the detectability of recombinational events.

Number of Recombinational Events in a Sample

One approach to estimating ρ attempts to infer the total number of recombination events in a sample of n sequenced alleles (R), and then relate this to the expectation for pairs of sites assumed to be in drift-mutation-recombination equilibrium. For neutral sites separated by L nucleotides, the expected value of R in a sample of n sequences is equal to $\rho L a_n$, where a_n has the same definition as given above for the expression of the number of segregating sites (Hudson and Kaplan 1985). Rearranging, a potential estimator for ρ is

$$\hat{\rho} = \hat{R}/(L a_n) \tag{4.6}$$

where \hat{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 4.3).

Given the invisibility of most recombinational events, the primary impediment to applying this expression is the estimation of R . As an entrée into this problem, Hudson and Kaplan (1985) proposed the four-gamete test, which asserts that any pair of 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 appropriately 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 nonparental haplotypes,

$$d_{r,\min} = 1 - [2\ln(1 + \Theta)]/\Theta + [1/(1 + \Theta)] \quad (4.7a)$$

$$d_{r,\max} = 1 - [2(1 - e^{-\Theta})]/\Theta + e^{-\Theta} \quad (4.7b)$$

where $\Theta = \theta L$ is the population mutation rate for the stretch of DNA being surveyed, with L being the length of the sequence (number of nucleotide sites). These two limits are respectively approached as $c \rightarrow 0.0$ (complete linkage) and 0.5 (free recombination among adjacent markers). 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 sites, the majority of recombination events will simply reproduce parental gamete types (Figure 4.2).

-Insert Figure 4.2 Here-

Given an estimate of Θ , the average result from Equations 4.7a,b can be used to estimate the total number of recombination events in the sample as R_{\min}/\bar{d}_r . However, this approach is not necessarily fully adequate because only a subset of the fraction of recombinant gametes that are nonparental with respect to markers are also novel with respect to the entire population, i.e., the fraction of 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 Lefebvre and Labuda 2008). Letting p_i denote the frequency of the i th haplotype in a sample, an estimator for the fraction of detectable (potentially informative) recombinant alleles is

$$\hat{d}_r = \sum_{\substack{i=1 \\ j>i}}^L 2p_i p_j L_{\max,ij} / L \quad (4.8)$$

where L is the maximum distance between segregating sites in the sample, and $L_{\max,ij}$ is the distance between the maximally separated heterozygous sites in the ij th comparison. Through simulations, one can establish the fraction of potentially

informative recombination events that do not lead to novel haplotypes in the sample, thereby reducing \hat{d}_r to d'_r , the fraction of recombination events that lead to novel recombinants. Recalling Equation 4.6, a method-of-moments estimator for the population recombination rate is then

$$\hat{\rho} = \hat{R}_{\min} / (L \hat{d}'_r a_n) \quad (4.9)$$

Other Approaches

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. In addition to an estimate of the average number of nucleotide differences between random sequences, $\theta_H L$, this approach requires only one other summary statistic, the observed variance of pairwise divergence,

$$\hat{\sigma}_k^2 = \frac{2}{n(n-1)} \sum_{\substack{i=1 \\ j>i}}^n (k_{ij} - \Theta_H)^2 \quad (4.10)$$

where k_{ij} is the number of sites at which sequences i and j differ. Wakeley's (1997) Equation 15 allows one to estimate ρL as a function of Θ_H , $\hat{\sigma}_k^2$, and n , and simple division by the length of the sequence (L) then yields $\hat{\rho}$.

A number of other approaches, conceptually developed in a maximum-likelihood (ML) framework, have been suggested. For example, focusing on a pair of informative segregating sites in a sample of four sequenced alleles, Hey and Wakeley (1997) found that under the assumptions of neutrality of markers and drift-mutation-recombination equilibrium, the probability of observing all four gametic types is

$$I \simeq \frac{2}{3} - \frac{1}{6} \left(\frac{\rho + 6}{\rho + 3} \right)^2 e^{-3\rho(L-1)/[5(\rho+3)]} \quad (4.11)$$

for $L \geq 1$, with a more complicated expression is given by the authors for $L = 0$ (adjacent sites). This statistic can be combined over multiple sample quartets and pairs of informative sites, solving for the estimate of ρ that maximizes the overall likelihood of the observed data. Strictly speaking, however, as some of the data will be nonindependent, this is not a conventional maximum-likelihood approach.

An alternative procedure makes fuller use of the data by evaluating the probabilities of various sample counts of the four gametic types at two loci (i.e., the numbers of **AB**, **Ab**, **aB**, and **ab** gametes) 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 be obtained by computer simulation. 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 is not equivalent 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 have been developed that 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 approaches. 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 and have very large sampling variances (Wall 2000).

However, as population-genomic data become widely available, the full power of ML approaches may eventually be realized. For example, Lynch (2008) has introduced an approach to obtaining large-scale estimates of ρ from a single randomly sampled diploid individual. This method uses the correlation of heterozygosity among pairs of sites separated by specific distances across the genome to obtain estimates of the squared disequilibrium coefficient $E(D^2)$ (Chapter 2) that are asymptotically unbiased with minimal sampling variance. In principle, with estimates of D^2 for neutral sites separated by 1, 2, 3, etc., nucleotides, each based on thousands to millions of pairs of sites, the decline in D^2 with L can be used to infer ρ with an expression such as Equation 2.26.

Empirical Observations

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

Species	ρ	c/u	References
Animals:			
<i>Drosophila melanogaster</i>	0.05846	3.545	Hey and Wakeley 1997 Andolfatto and Przeworski 2000
<i>Drosophila pseudoobscura</i>	0.08655	1.360	Hey and Wakeley 1997
<i>Drosophila simulans</i>	0.09720	3.306	Andolfatto and Przeworski 2000
<i>Homo sapiens</i>	0.00060	0.770	Frisse et al. 2001; Ptak et al. 2004 Lefebvre and Labuda 2008
Land plants:			
<i>Arabidopsis thaliana</i>	0.00160	0.193	Kim et al. 2007
<i>Brassica nigra</i>	0.00602	0.330	Lagercrantz et al. 2002
<i>Cryptomeria japonica</i>	0.00046	0.118	Fujimoto et al. 2008
<i>Helianthus annuus</i>	0.05280	4.100	Liu and Burke 2006
<i>Hordeum vulgare</i>	0.00080	1.417	Morrell et al. 2006
<i>Oryza rufipogon</i>	0.00003	0.006	Mather et al. 2007
<i>Oryza sativa</i>	0.00004	0.021	Mather et al. 2007
<i>Persea americana</i>	0.00338	0.582	Chen et al. 2008
<i>Pinus sylvestris</i>	0.01452	2.855	Pyhäjärvi et al. 2007
<i>Pinus taeda</i>	0.00175	0.266	Brown et al. 2004
<i>Solanum chilense</i>	0.02380	1.122	Arunyawat et al. 2007
<i>Solanum peruvianum</i>	0.03480	1.392	Arunyawat et al. 2007
<i>Sorghum bicolor</i>	0.00041	0.130	Hamblin et al. 2005
<i>Zea mays</i>	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 effects of the long-term effects of selection, certainly not selective sweeps or background selection on linked chromosomal regions. Nevertheless, given

the effects that N_e 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 the value of N_e that best explains the data (reviewed by Wang 2005). From the standpoint of natural populations, only one approach appears to harbor much promise – monitoring temporal changes in putatively neutral allele frequencies within an isolated population and back-calculating the value of N_e that best accounts for the observed fluctuations.

Consider a single polymorphic locus 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 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-level 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 expected estimate of the 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 \hat{p}_0 and \hat{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 $E[(\hat{p}_1 - \hat{p}_0)^2]$ because $E(\hat{p}_1 - \hat{p}_0) = 0$ 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 as

$$\hat{N}_e = \frac{1}{2\hat{F}_1 - (1/n_0) - (1/n_1)} \quad (4.12a)$$

where

$$\hat{F}_1 = \frac{(\hat{p}_0 - \hat{p}_1)^2}{\hat{p}_0(1 - \hat{p}_0)} \quad (4.12b)$$

is a measure of the standardized variance of allele-frequency change. With multiple loci, the estimate \hat{F} is obtained by averaging over loci. 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 4.12a.

Despite their intuitive nature, Equations 4.12a,b yield biased estimates because the contributions of the sampling variance 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). In addition, \hat{F}_1 is undefined if $\hat{p}_0 = 0$, and Equations 4.12a,b do not immediately allow for the incorporation of multiple alleles ($k > 2$). An alternative estimator that deals with these problems is

$$\hat{N}_e = \frac{t-2}{2\hat{F} - (1/n_0) - (1/n_1)} \quad (4.13a)$$

where \hat{F} is calculated by either

$$\hat{F}_2 = \frac{1}{k} \sum_{i=1}^k \frac{(\hat{p}_{0i} - \hat{p}_{1i})^2}{[(\hat{p}_{0i} + \hat{p}_{1i})/2] - \hat{p}_{0i}\hat{p}_{1i}} \quad (4.13b)$$

(Nei and Tajima 1981b), or

$$\widehat{F}_3 = \frac{1}{k} \sum_{i=1}^k \frac{(\widehat{p}_{0i} - \widehat{p}_{1i})^2}{(\widehat{p}_{0i} + \widehat{p}_{1i})/2} \quad (4.13c)$$

(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 generally lead to very similar results (Waples 1989). One drawback of Equation 4.13a is that it requires that the generation interval exceed two.

As noted above, more refined estimates of F can be obtained by averaging estimates of F_2 or F_3 over multiple loci, and Pollak (1983) derived a more generalized estimator that allows for sampling across more than a single 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, where 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.

Regardless of the method used, it should be noted that 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

$$\text{Var}(\widehat{N}_e) \simeq \left(\frac{8N_e^4}{t^2M} \right) \left(\frac{1}{4N_e^2} + \frac{1}{N_e t \bar{n}} + \frac{1}{t^2 \bar{n}^2} \right) \quad (4.14)$$

where M denotes the number of independent allelic comparisons (the sum of $k - 1$ over all loci) and \bar{n} is the harmonic mean of the sample sizes in the two generations (Pollak 1983). In general, M , t , and \bar{n} will be under the control of the investigator, so the form of Equation 4.14 provides a useful basis for designing an optimal sampling strategy. For example, a doubling of M will always 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. If $N_e \gg t\bar{n}$, the middle term on the right becomes negligible, showing that a doubling of \bar{n} will also have a much greater effect than a doubling in M . The sampling distributions of $M\widehat{F}/E(F)$ are χ^2 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 4.13a, e.g., using the values of F at the 2.5 and 97.5% cumulative probability levels to yield 95% confidence limits.

As noted in previous contexts, because of their simple heuristic interpretation, methods-of-moments estimators, like those just noted, are highly popular approaches

to estimating population parameters. However, by simply using 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 capable of yielding estimates that best explain the entire distributions 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) and Anderson (2005) present approximations that are computationally efficient.

Empirical Observations

In Chapter 3, we encountered numerous demographic factors that influence the effective size of a population, almost always 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 that average N_e/N is 11%, whereas a subsequent study with a much larger sample obtained an average of 14% (Palstra and Ruzzante 2008).

It is likely that the $\sim 90\%$ reduction in N_e suggested by the preceding results is a considerable underestimate of the situation for many nonvertebrate species and even many vertebrates when long-term considerations are taken into account. For example, high-fecundity fish in spatially variable environments appear to have $N_e/N < 0.001$ (Hedrick 2005). In addition, almost all existing studies have involved obligate outcrossers, and because many unicellular species have conspicuous phases of asexual reproduction that can encourage the rapid proliferation of a small number of clones, N_e/N ratios much lower than 0.001 are likely to be quite common in such species (see below). Indeed, 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.

MUTATION RATE

Because mutations arise at an extremely low rate per nucleotide site, estimation of the rate of mutation is formidably challenging, with most approaches relying on procedures that attempt to enrich the pool of experimentally derived mutations in an effectively neutral fashion. Here, we review the two most commonly used methods of enrichment: 1) long-term genome-wide accumulation of mutations in isolated lineages in an effectively neutral fashion; and 2) short-term isolation of conspicuous mutants at single loci from a large population raised on a selective medium.

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 homozygous 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, and then the progeny of one parent can be used to synchronously initiate a set of parallel lines, each to be subsequently maintained by single-progeny descent. Under this design, essentially all mutations that do not cause lethality or complete sterility will accumulate independently in each line at a rate of u per generation per nucleotide site 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 dozen to hundreds of generations of mutation accumulation nearly all mutations can be detected as fixed homozygotes by a sequencing a subset of lines. Typically, nearly all lines will be identical at individual nucleotide sites (reflecting the ancestral state), whereas mutations will appear 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, u is estimated as $\hat{u} = m/(nLT)$. The sampling variance of \hat{u} using this approach is $\sim u/(nLT)$, yielding a coefficient of variation of $(unLT)^{-1/2}$, 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 divergence of random alleles within a species has expected value $4N_e u$. 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 $d = 2tu + 4N_e u$, assuming equal N_e in both taxa. At $t = 0$, $d = 4N_e u$ (the average divergence of randomly sampled alleles in the ancestral population), whereas as $t \rightarrow \infty$, $d \simeq 2tu$ (a widely used approximation in applications of molecular clocks for dating evolutionary events). Rearranging, and letting $\bar{\theta}_H$ denote the average within-species nucleotide diversity at silent sites, we obtain an estimator for the mutation rate, $\hat{u} = (\hat{d} - \bar{\theta}_H)/(2t)$.

To procure an estimate of the mutation rate for humans, Nachman and Crowell (2000) obtained the parallel sequences of 12 unexpressed pseudogenes in human and chimpanzee. Because they are not expressed, such stretches of DNA are expected to fulfill the assumptions of neutrality. The average number of substitutions per site separating the two species was 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), yielding an average of $\bar{\theta}_H = 0.0011$. Nachman and Crowell (2000) assumed a divergence time of 5 million

years, and an average generation time of 20 years, yielding $t \simeq 250,000$. Substitution into the preceding expression then gives an estimated mutation rate of 2.44×10^{-8} per site per generation for base substitutions.

Short-term Enrichment

The preceding approach emphasizes a strategy of augmenting the pool of mutations by passing a set of lines through a large number of generations. The advantage of such a protocol is that mutations will be equally enriched throughout the genome, minimizing the chances that the mutational profile will be biased by observations at any particular target locus (assuming results are procured over many loci). An alternative approach, widely applied to microbial cultures, is to focus on visible markers (those causing obvious phenotypic changes) associated with mutations at particular loci. Here the emphasis is on isolating neutrally accumulated mutations out of a very large pool of cells in a relatively short period of time, 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 permits growth on a selective medium (Luria and Delbrück 1943). Given estimates of the total number of cells in the culture, and the number of these that are mutant, it is then possible to estimate 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 (r) to the number of mutations leading to them (m). In addition, not all mutations produce an observed phenotype, so the second challenge is to determine the fraction of mutations at the target locus that are detectable (d). The true number of mutations is equal to m/d . Finally, in order to determine the mutation rate per nucleotide site, one must know the mutational target size (n). Thus, for this 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 (Rosche and Foster 2000), with broad technical overviews being provided by 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

$$E(p_0) = e^{-m} \quad (4.15a)$$

Rearrangement leads to the estimator $\hat{m} = -\ln(p_0)$, 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 cannot be procured 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 it completely ignores the distribution of mutant numbers in different replicate cultures.

Full use of such information can be achieved by relying on a maximum-likelihood framework. From Lea and Coulson (1949) and Sarkar et al. (1992), conditional on m being ≥ 1 , the expected frequency of cultures containing r mutant cells is

$$p_r(m) = \frac{m}{r} \sum_{i=0}^{r-1} \frac{p_i(m)}{r-i+1} \quad (4.15b)$$

provided the starting population size is much smaller than that at the end point. This expression can be solved recursively starting with Equation 4.15a for $p_0(m)$, and iterating up to the maximum observed level of r . Letting L_r be the number of cultures observed to have r mutant cells, the log likelihood of m given the entire pool of data is

$$L(m) = \sum_{r=0}^{r_{\max}} L_r \ln[p_r(m)] \quad (4.16)$$

The ML estimate is the value of m that maximizes $L(m)$.

Example 4.2. An alternative approach to estimating the mutation rate in an exponentially growing culture is to consider the dynamics of the change in the frequency of mutant cells in the population. Letting f_0 be the initial frequency of mutations, ϕ be the rate of exponential growth of the cell culture (assumed to be identical for mutant and nonmutant cells), and u_o be the rate of mutation to an observable phenotype per cell division, the expected frequency after t time units is

$$f_t = f_0 + (1 - f_0)(1 - e^{-u_o \phi t})$$

This follows from the fact that $e^{-u_o \phi t}$ is the probability that a descendent cell has not acquired a detectable mutation after ϕt cell divisions. Note that if one starts with a mutation-free culture ($f_0 = 0$) and the cumulative probabilities of mutation ($u_o \phi t$) are $\ll 1$, the expected fraction of mutant cells will increase in an essentially linear fashion, at rate $u_o \phi$. Viewing ϕt as the average number of divisions experienced by a cell through its ancestral lineage, the regression slope of a plot of f_t on ϕt then provides an estimate of u_o .

Thus, with this approach, the mutation rate per cell division can be estimated from the cumulative behavior of a single expanding culture. However, because of the stochastic nature of mutations, the results from single cultures are not terribly reliable. 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 and survey the frequency of mutants at the end point of each culture. Rearrangement of the preceding expression yields the relevant point estimator of the mutation rate to observable phenotypes,

$$\hat{u}_o = -\frac{\ln[(1 - f_0)/(1 - f_t)]}{\phi t}$$

Because $N_t = N_0 e^{\phi t}$ 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 , this expression further simplifies to

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

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

Drake (1991) argues 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. 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

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

Given just the number of cells and the frequency of mutants at the end point, this expression can be solved recursively to obtain the estimate \hat{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 actual mutation rate requires that the fraction of mutations that are detectable at the marker locus (d) be known. At a minimum, this requires that the target locus of a large number of independent mutant cells be sequenced to ascertain their molecular basis. Generally, because the mutation rate per nucleotide site is quite low, no more than a single change will be found within a sequenced locus, and there is little ambiguity in the identity of the causal mutation. For base-substitutional mutations, Drake (1991) makes the following argument for obtaining an estimate of d . Assuming that all mutations causing premature termination of translation (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 population (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), $d = 3n_o/(64n_n)$ is an estimate of the fraction of total base-substitutional mutations that are detectable (if all detected base-substitutional mutations were to termination codons, $n_n/n_o = 1$, and $d = 3/64$). If n is the length of the target sequence over which mutations are detectable (generally assumed to be the length of the coding region, which could be an overestimate), an estimator of the base-substitutional rate per nucleotide site is

$$\hat{u} = \frac{\hat{u}_o}{dn}$$

Empirical Observations

Although accurate estimates of the mutation rate are still 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 0.83 and 0.72, respectively, for reporter-construct studies and the complete sequencing of mutation-accumulation lines of the yeast *Saccharomyces cerevisiae* (Lynch 2006; Lynch et al. 2008). On a per-generation basis, they are 5.68 and 9.45, respectively, for sequenced mutation-accumulation lines of the fly *Drosophila melanogaster* (Haag-Liautard et al. 2007) and the nematode *Caenorhabditis elegans* (Denver et al. 2004), and averaging the results from Examples 4.1 and 4.3, the rate for humans is 25.64×10^{-9} per site per generation. The average base-substitutional mutation rate for ten prokaryotic species is 0.88×10^{-9} per site per cell division (Lynch 2006, and unpubl. data), similar to that for unicellular yeast.

These limited results strongly suggest that the per-generation rate of mutation is substantially increased in multicellular relative to unicellular species. Although the mechanisms underlying this inflation are unknown, an obvious distinction between these two groups is that the former experience multiple germline cell divisions per generation – averaging over the two sexes, ~ 10 for *C. elegans*, 36 for *D. melanogaster*, and 200 for *H. sapiens* (Drost and Lee 1995; Kimble and Ward 1998; Crow 2000). Thus, in principle, the elevated mutation rate in multicellular species may be associated with replication-dependent mutations, although the unicellular : invertebrate : human ratio of the per-generation rates (1 : 10 : 34) is considerably smaller than the ratio of the number of cell divisions/generation (1 : 23 : 200). Because the average amount of time between cell divisions is ~ 0.2 days in yeast, ~ 0.4 days in flies and nematodes, but on the order of a month in humans, the observed scaling may also in part be associated with replication-independent mutations.

These observations suggest that the base-substitutional mutation rate in multicellular species may not be elevated either on the scale of cell divisions or absolute time. However, for a rather different class of sites, dinucleotide repeats contained within microsatellite loci, the ratio of per-cell division mutation rates to repeat-number variants is 1 : 5 : 67 for yeast, *C. elegans*, and humans (Seyfert et al. 2008), implying a substantial increase in mutagenicity with multicellularity. Such observations are in accordance with empirical evidence that the basic DNA repair-pathway machinery has evolved to different levels of efficiency in various phylogenetic lineages (Lynch 2008), implying that differences in mutation rates cannot be totally explained on the basis of cell-division numbers and generation lengths. Variation in the spectrum of mutational effects also exist, with the ratio of insertion/deletion to base-substitutional mutations being ~ 1.0 in *C. elegans*, but in the range of 0.10 to 0.25 in yeast, *Drosophila*, and humans (Lynch et al. 2008).

Example 4.3. To indirectly estimate the human mutation rate, Kondrashov (2002) took advantage of records on the incidence of genetic pathologies attributable to dominant mutations at known causal loci. The population frequency of genetic disorders (I) 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 equal to $p = u/s$, where u is the

mutation rate to defective alleles (per gene copy), and s is the selective disadvantage of affected individuals (see below, Equation 7.3). 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, and for this we will employ a strategy similar in spirit to that advocated by Drake (1991). For each disorder in the survey of Kondrashov (2002), 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 mutation. Assuming all mutations involving insertions and deletions were detectable, from the incidence of chain-terminating base-substitutional mutations, the detectability of mutations can then be calculated, as outlined in Example 4.2.

Although Kondrashov's (2002) survey involved 32 different genetic disorders, we will simply follow the steps involved in the calculations for one such analysis, following this by a summary of the total pool of 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 to be 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, 88% of which involves base-substitutional mutations.

When these approaches are extended to the remaining 31 loci, the estimated average total mutation rate per site is 2.87×10^{-8} , whereas that for base-substitutional changes alone is 2.64×10^{-8} . Note that this latter estimate is remarkably close to that reported in Example 4.1, derived from long-term interspecies divergence estimates.

RECOMBINATION RATE

Although it is extraordinarily difficult to estimate recombination rates at individual 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 observations on rates of meiotic crossing-over between informative molecular markers, now available for well over 100 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, an enterprise made difficult by the fact that only odd numbers of crossovers between markers lead to recombinant genotypes (LW Chapter 14). Lengths of chromosomes 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 in 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). There appears to be a simple physical explanation for such behavior. During meiosis, there are typically no more than two crossover events per chromosome, so that average chromosome lengths are generally in the range of 0.5 to 2.0 Morgans, regardless of chromosome size. Thus, because phylogenetic increases in genome sizes are generally caused by 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, N be the number of chromosomes in a haploid set, and x be the average number of crossovers (Morgans) per chromosome, then $\bar{c} \simeq xN/G$, assuming that x is independent of chromosome size. If this model is correct, then 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 by variation in xN . The data closely adhere to this predicted pattern, with the smallest genomes of microbial eukaryotes exhibiting 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 4.3). Over this entire gradient, there is a smooth, overlapping decline in recombination intensity across unicellular species, invertebrates, vertebrates, and land plants, reflecting the general increase in genome sizes between these eukaryotic domains.

Thus, the vast majority of the variance in recombination rate among eukaryotic species is due to variation in genome size and chromosome number. Remarkably, using approaches to be outlined in Chapter 5, Dumont and Payseur (2007) find that variation in recombination rates across mammalian species evolves in a manner that cannot be discriminated from the expectations under a neutral model.

It should be noted that even in the highest density genetic maps, adjacent markers are generally separated by tens of thousands to millions of base pairs, so the kinds of results summarized above only yield measures of average levels of recombination at the genomic level. Considerable heterogeneity in recombination rates, up to 100-fold differences, can exist within chromosomes, with highly localized recombinational hot spots existing in some species (Petes 2001; de Massy 2003; Jeffreys et al. 2004; Myers et al. 2005; Arnheim et al. 2007). Nevertheless, the

results in Figure 4.3 imply that, depending on the species, chromosomal segments of 10^4 to 10^6 nucleotide sites may remain intact through hundreds to thousands of meioses. In humans, for example, average c is $\sim 10^{-8}$ per site per generation, which implies that a typical ~ 10 -kb stretch of human DNA has just a 10% chance of experiencing a recombination event over a 1000-generation period. Thus, it is not surprising that surveys of molecular markers reveal many large (10 to 100-kb) regions in the human genome that appear to have experienced little internal recombination (Daly et al. 2002).

Table 4.3. Basic features of the physical and genetic maps of various eukaryotic groupings, derived from a large survey of mapping studies involving high-density molecular markers. The grouping “Other unicellular species” include 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

GENERAL IMPLICATIONS

Although limited in many respects, the results summarized above allow us to make several generalizations that will prove useful in subsequent chapters. First, given an estimate of $\theta = 4N_e u$ and u , it is possible to estimate the long-term effective population size of a species by factoring out the latter. Noting that the average estimate of θ for unicellular eukaryotes is 0.057 (which is likely somewhat downwardly biased), and that u for base-substitutional mutations in such species is $\sim 0.8 \times 10^{-9}$, the average N_e for such species appears to be on the order of 4×10^7 individuals if haploidy is assumed (and half that if diploidy is assumed). Using an average θ of 0.026 and u of 7.6×10^{-9} for invertebrates, average N_e for this grouping is $\sim 10^6$, and using $\theta = 0.0011$ (Example 4.1) and $u = 25.6 \times 10^{-9}$, N_e for the human population is $\sim 10,000$.

Similar indirect inferences can be made from estimates of $\rho = 4N_e c$. For example, from Table 4.1, the average estimate of ρ for *Drosophila* species is 0.0807, whereas ρ for humans is ~ 0.0006 , and for annual plants and trees is 0.0134 and 0.0050, respectively. From the genetic map data contributing to Figure 4.3, average c ($\times 10^{-8}$ per site per generation) is 2.14 for *Drosophila*, 1.28 for humans, 1.59 annual plants,

and 2.93 for trees. These results imply average values of N_e of $\sim 10^6$ for *Drosophila*, 10,000 for humans, 200,000 for annual plants, and 40,000 for trees. The consistency of the results from both approaches when applied to both *Drosophila* and humans is compelling.

Because the mean coalescent time for a random pair of alleles is $2N_e$ generations (Chapter 2), the preceding indirect estimates of N_e are expected to be reasonable approximations of the average conditions experienced over periods of $\sim 2N_e$ generations. Moreover, these estimates should be considered simply as broad indicators, as θ (and therefore N_e) can vary by an order of magnitude within major phylogenetic groups (Lynch 2006). Nevertheless, the overall picture that emerges is that long-term effective population sizes are typically orders of magnitudes smaller than the actual numbers of breeding adults within species (even more so than implied by the short-term estimates of N_e given above for vertebrates). As a likely consequence of the effects of selection on mutations physically linked on chromosomes (Chapter 3), the upper limit to N_e is probably not much beyond 10^9 , even in the most enormous microbial populations.

Taken together, the above results lead to several general conclusions about the relative power of the three nonadaptive forces of evolution and the manner in which they scale across various phylogenetic groups. First, mutation rates increase from $\sim 10^{-9}$ per nucleotide site per generation in unicellular species to between 10^{-8} and 10^{-7} in long-lived, multicellular species. Second, average genome-wide recombination rates scale positively with the number of chromosomes but inversely with total genome size in a manner that is independent of the of the number of germline cell divisions per genome (because single meiotic events lead to the production of each newborn). Third, except in highly self-fertilizing species, the recombination rate per nucleotide site is within a factor of four of the mutation rate. Ratios of c/u on the order of 1.0 and smaller imply a significant vulnerability to linkage effects resulting from the inability of recombination to separate independently arising mutations, with c/u ratios between selected sites greater than four or so being necessary to completely alleviate problems with Hill-Robertson interference (Marais and Piganeau 2002). Thus, there may be no general selective advantage to increase c beyond a level of $\sim 4u$. Fourth, the power of random genetic drift generally substantially exceeds the power of both mutation and recombination at the level of nucleotide sites, almost always by at least one to two orders of magnitude (more so in large multicellular species). Fifth, at least for species with very large numbers of reproductive adults, the power of random genetic drift appears to be much more strongly governed by the effects of genetic hitch-hiking and background selection rather than by simple stochastic effects of gamete sampling. This contention is consistent with the low ratios of c/u that are universally observed across species, which encourage such effects.

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. Przeworski. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* 156: 257-268. [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]
- 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]
- 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]
- 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]
- Crow, J. F. 2000. The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics* 1: 40-47. [4]
- Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. 2002. High-resolution haplotype structure in the human genome. *Nature Genetics* 29: 229-232. [4]
- de Massy, B. 2003. Distribution of meiotic recombination sites. *Trends Genet.* 19: 514-522. [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]
- Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA* 88: 7160-7164. [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]
- 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]
- 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]
- 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. Popul. Biol.* 48: 172-197. [4]
- Fu, Y.-X., and W.-H. Li. 1993. Maximum likelihood estimation of population parameters. *Genetics* 134: 1261-1270. [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.* (in press). [4]
- Golding, G. B. 1984. The sampling distribution of linkage disequilibrium. *Genetics* 108: 257-274. [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]
- 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]
- Hedrick, P. 2005. Large variance in reproductive success and the N_e/N ratio. *Evolution* 59: 1596-1599. [4]
- Hey, J., and J. Wakeley. 1997. A coalescent estimator of the population recombination rate. *Genetics* 145: 833-846. [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]
- 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]
- Johnson, P. L., and M. Slatkin. 2008. Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25: 199-206. [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]
- Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* 21: 12-27. [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., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* 156: 1393-1401. [4]
- Lagercrantz, U., M. Kruskopf 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]
- 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]
- 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]
- 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. 2008. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* (in press). [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]
- Marais, G., and G. Piganeau. 2002. Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. *Mol. Biol. Evol.* 19: 1399-1406. [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]
- 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]

- 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., 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]
- Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154: 931-942. [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.* (in press). [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. *Nature Reviews 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]
- 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]
- Rosche, W. A., and P. L. Foster. 2000. Determining mutation rates in bacterial populations. *Methods* 20: 4-17. [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]
- 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]

- 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]
- 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]
- Stephens, J. C. 1986. On the frequency of undetectable recombination events. *Genetics* 112: 923-926. [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]
- 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. Popul. Biol.* 7: 256-276. [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]
- 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., V. Yotova, D. Gehl, T. Wambach, I. Arrieta, M. Batzer, D. E. Cole, P. Hechtman, F. Kaplan, D. Modiano, J. P. Moisan, R. Michalski, and D. Labuda. 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, 4.4, and 4.5, for $\theta = 0.1, 0.01, \text{ and } 0.001$ in descending order. The assumed number of sites is $L = 10,000$ in all cases.

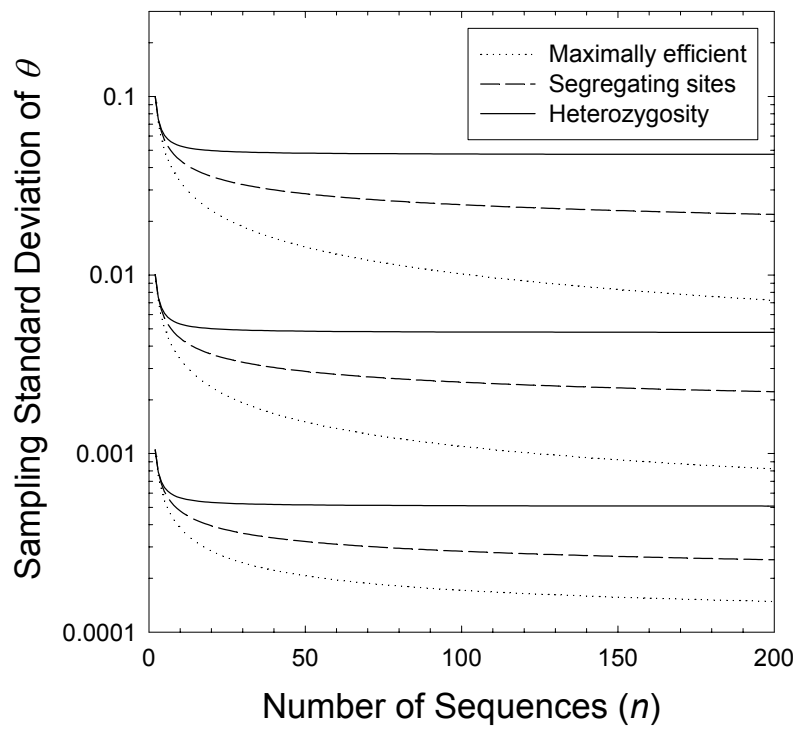


Figure 4.2. The upper and lower bounds on the fraction of recombination events that produce nonparental gametes among one or more neutrally evolving sites (from Equations 4.7a,b). Θ is the product of the population mutation rate per site (θ) and the length of the segment (in base pairs).

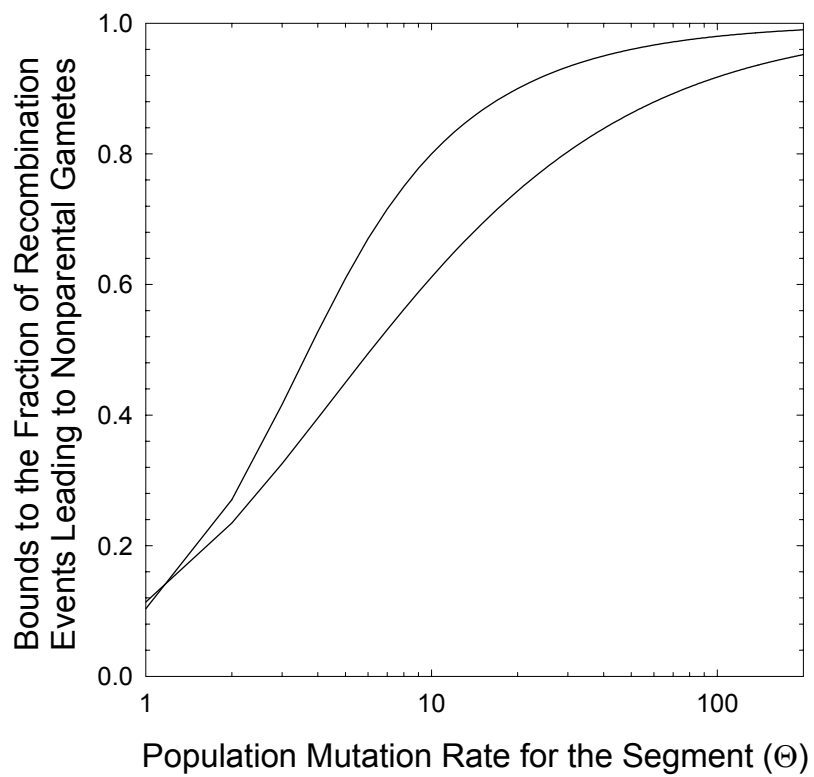


Figure 4.3. Average rates of recombination per physical distance for four major groupings of eukaryotes, determined from information on total physical and genetic map sizes. The two dashed lines have slopes of -1.0 in accordance with the theory discussed in the text. The top line assumes $xN = 50$, i.e., 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, average x is in the range of 0.3 to 3.1 (with one exception) and N is in the range of 3 to 44.

