Population Genetics of Polymorphism and Divergence

Stanley A. Sawyer[*,†] **and Daniel L. Hartl**[†]

[*]*Department of Mathematics, Washington University, St. Louis, Missouri 63130,* [†]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110*

November 3, 1992

ABSTRACT

Frequencies of mutant sites are modeled as a Poisson random field in two species that share a sufficiently recent common ancestor. The selective effect of the new alleles can be favorable, neutral, or detrimental. The model is applied to the sample configurations of nucleotides in the alcohol dehydrogenase gene *(Adh)* in *Drosophila simulans* and *D. yakuba* (McDonald and Kreitman 1991, *Nature* **351:** 652–654). Assuming a synonymous mutation rate of 1.5×10^{-8} per site per year and 10 generations per year, we obtain estimates for the effective population size ($N_e = 6.5 \times 10^6$), the species divergence time ($t_{div} = 3.74$ Myr), and an average selection coefficient ($\sigma = 1.53 \times 10^{-6}$ per generation for advantageous or mildly detrimental replacements), although it is conceivable that only two of the amino acid replacements were selected and the rest neutral. The analysis, which includes a sampling theory for the independent infinite sites model with selection, also suggests the estimate that the number of amino acids in the enzyme that are susceptible to favorable mutation is in the range 2–23 out of 257 total possible codon positions at any one time. The approach provides a theoretical basis for the use of a 2×2 contingency table to compare fixed differences and polymorphic sites with silent sites and amino acid replacements.

It has been more than 25 years since Lewontin and Hubby (1966) first demonstrated high levels of molecular polymorphism in *Drosophila pseudoobscura*. This finding had two strong immediate effects on evolutionary genetics: it stimulated molecular studies of many other organisms, and it led to a vigorous theoretical debate about the significance of the observed polymorphisms (Lewontin 1991). The experimental studies soon came to a consensus in demonstrating widespread molecular polymorphism in numerous species of plants, animals, and microorganisms. The theoretical debate was not so quickly resolved. One viewpoint (Kimura 1968, 1983) held that most observed molecular variation within and among species is essentially selectively neutral, with at most negligible effects on survival and reproduction. Opposed was the classical Darwinian view that molecular polymorphism is the raw material from which natural selection fashions evolutionary progress, and that the newly observed molecular variation was unlikely to be any different (Lewontin 1974). The two viewpoints could not have been more at odds, and a great controversy ensued. To a large extent the issue has been clouded by inadequate data (Lewontin 1974, 1991). Observations of natural populations are snapshots of particular places and times, and the resulting inferences about the long-term fate of molecular polymorphisms can be challenged by neutralists and selectionists alike. By the same token, laboratory experiments capable of detecting selection coefficients as small as are likely to be important in nature are currently impractical (Hartl and Dykhuizen 1981, Hartl 1989), although some large effects have been documented (see Powers *et al.* 1991 for a review).

In the 1980s, the increasing use of DNA sequencing in evolutionary genetics gave some hope that the impasse could be overcome. Direct examination of genes, rather than the electrophoretic mobility of gene products, yields vast amounts of information consisting of hundreds or thousands of nucleotides. The data are also of a different quality, since the DNA sequences are unambiguous and

contain both synonymous nucleotide differences and differences that change amino acids. To the extent that the synonymous differences are subjected to weaker selective effects than amino acid differences, comparisons between the two types of polymorphisms can serve as a basis of inference. Synonymous polymorphisms are more common than amino acid polymorphisms (KREITMAN 1983), and also appear to be more weakly affected by selection (SAWYER, DYKHUIZEN, and HARTL 1987).

With data from only one species, the level of synonymous and replacement polymorphism must be substantial in order for statistical analysis to have enough power to detect selection (SAWYER, DYKHUIZEN, and HARTL 1987; HARTL and SAWYER 1991). Most eukaryotic genes are not sufficiently polymorphic to allow this approach. An alternative approach, pioneered by HUDSON, KREITMAN, and AGUADÉ (1987), is based on comparing polymorphisms within species with fixed differences between species. This approach has been applied to the Drosophila fourth chromosome (BERRY, AJIOKA, and KREITMAN 1991) as well as to the tip of the X chromosome (BEGUN and AQUADRO 1991), both of which are regions of reduced recombination. The level of polymorphism in these regions is also reduced, and the analysis suggests strongly that the reduction is the result of genetic hitchhiking associated with periodic selective fixations.

Comparison of molecular variation within and between species is also the crux of a statistical test proposed by MCDONALD and KREITMAN (1991a). The test is for homogeneity of entries in a 2×2 contingency table based on aligned DNA sequences. The rows in the contingency table are the numbers of replacement or synonymous nucleotide differences, and the columns are either the numbers of fixed differences between species or else of polymorphic sites within species. Here *polymorphic sites* are defined as sites that are polymorphic within one or more of the species, and *fixed differences* are defined as sites that are monomorphic (fixed) within each species but differ between species. The term *silent* refers to nucleotide differences in codons that do not alter the amino acid, and *replacement* refers to nucleotide differences within codons that do alter the amino acid. The McDonald-Kreitman test compares the number of silent and replacement polymorphic sites with the number of silent and replacement fixed differences. When 30 aligned DNA sequences from the alcohol dehydrogenase *(Adh)* locus of three species of *Drosophila* were compared (MCDONALD and KREITMAN 1991a), there were too few polymorphic replacement sites ($P = 0.007$, two-sided Fisher exact test). MCDONALD and KREITMAN (1991a) argue that the most likely reason for the discrepancy is that some of the amino acid differences were fixed as a result of positive selection acting on replacement mutations. The possibility that the fixed differences could have resulted from a combination of slightly deleterious alleles (OHTA 1973), coupled with a dramatically changing population size, was also considered by MCDONALD and KREITMAN (1991a) but considered implausible because this would seem to require extraordinarily fine tuning among a large number of independent parameters.

Although the McDonald-Kreitman test has considerable intuitive appeal, little quantitative theory exists for the comparison of intraspecific polymorphism with interspecific divergence in the presence of selection. In this paper we present such a theory. Among other things, it addresses the question of whether the imbalance in the *Adh* contingency table could have resulted from the random fixation of mildly deleterious alleles over an extremely long time in a population of constant size, rather than fixations of advantageous alleles in a shorter period of time. The theory also provides an estimate of the average amount of selection required to produce the discrepancy observed, as well as an estimate of the rate at which favorable mutations occur (or, equivalently, an estimate of the average number of amino acids in the protein that are susceptible to a favorable mutation at any one time). Several objections to the details of the implementation of the McDonald-Kreitman test have been raised (GRAUR and LI 1991, WHITTAM and NEI 1991), and these are also addressed briefly.

Rationale and Results of the Analysis

The first step in our method is to analyze the sample configurations of the nucleotides occurring at synonymous sites in the aligned DNA sequences under the assumption that the synonymous variation is selectively neutral. This information is used to estimate the mutation rate at silent sites and the

divergence time between pairs of species. The divergence time is critical because, if the divergence time between species is sufficiently long, then conceivably all of the fixed amino acid differences between species could be due to the fixation of mildly deleterious alleles, and the significance of the McDonald-Kreitman contingency table might be an artifact of saturation at silent sites. Using the estimated values of the silent mutation rate and the divergence time, the numbers of synonymous polymorphic sites and fixed differences predicted from the neutral configuration theory are compared with the observed numbers. These estimates fit the observed *Adh* data very closely for all three pairwise species comparisons, which suggests that the configuration distributions at synonymous sites are roughly consistent with an equilibrium neutral model.

The second step is to develop equations for the expected number of polymorphic sites and fixed differences between a pair of species in terms of the magnitude and direction of selection, the mutation rate to new alleles having a given (constant) selective effect, and the divergence time. From these equations we estimate the amount of selection needed to explain the observed deficiency or excess in the number of replacement polymorphisms. We also estimate the rate of new mutations resulting in amino acid replacements (or, equivalently, the number of amino acid sites in the protein product at which favorable or mildly deleterious substitutions are possible at any one time). While the configuration analysis takes into account the possibility of multiple mutations at the same site, the second step assumes that this does not occur; i.e., that the genetic locus involved has not been saturated by mutations since the divergence of the two species. This assumption was checked in two different ways. First, the expected number of silent polymorphisms was calculated by both methods and found to agree within 12%. Second, the expected number of synonymous sites with two or more neutral fixations since the species diverged was estimated as less than two in all cases. There are no silent site polymorphisms with three or more nucleotides in the data considered, and only one silent site (between *D. melanogaster* and *yakuba*) is polymorphic in both species.

Most of our analysis depends on the assumption of linkage equilibrium or independence between sites. There is considerable linkage disequilibrium in *Adh* around the *Fast* versus *Slow* electrophoretic polymorphism in *D. melanogaster* (but not in *D. simulans* and *D. yakuba*). Possible balancing or clinal selection on this polymorphism may not only affect nucleotide configurations in *D. melanogaster*, but also may not be appropriate for the model of genic selection that we apply below. Among the three species for which McDonald and Kreitman (1991a) have *Adh* sequences, we are most confident in applying the analysis to the *D. simulans* versus *D. yakuba* comparison. The resulting analysis of the joint nucleotide configurations at silent sites for *D. simulans* and *D. yakuba* leads to the following estimates for the scaled silent mutation rate μ_s (summed over synonymous sites) and the species divergence time t_{div}:

$$\mu_s = 2.05 \quad \text{and} \quad t_{div} = 5.8 \tag{1}$$

both scaled in terms of the haploid effective population size N_e. That is, $\mu_s = u_s \times N_e$, where u_s is the nucleotide mutation rate per generation summed over all synonymous sites within amino acid monomorphic codon positions in *Adh* other than those coding for leucine and arginine (i.e., at "regular" silent sites; see below). Analysis of the numbers of replacement polymorphisms and fixed differences then leads to the following minimal estimates for the effective aggregate replacement mutation rate μ_r and average selection coefficient γ:

$$\mu_r = 0.013\mu_s \quad \text{and} \quad \gamma = 9.95 \tag{2}$$

again scaled in terms of the haploid effective population size. The quantity μ_r is the aggregate base mutation rate causing advantageous or mildly deleterious amino acid changes. The estimate of μ_r in (2) includes a correction factor to account for silent polymorphisms that are destined to be fixed but are not yet fixed. The quantity γ is the estimated average selection coefficient among advantageous or mildly deleterious amino-acid changing mutations, where "average" here means that γ is the selection coefficient required to produce the same numbers of replacement polymorphisms and fixed differences if these replacement mutations all had the same selective effect. The quantity γ is scaled in terms of the effective population size; i.e., $\gamma = \sigma N_e$, where σ is the same selection coefficient per generation. In

this case, the estimates of γ and σ are the minimum amount of selection required. Since there are no replacement polymorphisms between *D. simulans* and *D. yakuba* in the McDonald-Kreitman data, any larger value than $\gamma = 9.95$ in (2), along with a correspondingly smaller value of μ_r, would explain the data just as well.

The estimated aggregate mutation rate at silent sites of $\mu_s = 2.05$ in (1) is based on 212 regular amino acid monomorphic codon positions (see below). This estimate therefore corresponds to $2.05/212 = 0.0097$ silent nucleotide changes per synonymous site per N_e generations in the *Adh* sequence. Assuming a silent nucleotide substitution rate of 0.015 per synonymous site per million years (Myr) at the *Adh* locus, estimated from data on Hawaiian *Drosophila* (Rowan and Hunt 1991), N_e generations is 0.645 Myr. Therefore, $t_{div} = 5.8$ in (1) implies a divergence time of 3.74 Myr between *D. simulans* and *D. yakuba*. This value is in the middle of a range 1.6–6.1 Myr implied by single-copy nuclear DNA hybridization data (Caccone, Amato, and Powell 1988), where an estimated divergence time between *D. melanogaster* and *D. simulans* of 0.8–3 Myr (Lemeunier *et al.* 1986) is used as the standard of comparison. As a consistency check, the same analysis was carried out with the *Adh* data from *D. simulans* and *D. melanogaster*. This analysis yielded $\mu_s = 2.07$ for 213 monomorphic codon positions and a value of $t_{div} = 1.24$, from which the estimated divergence time is 0.80 Myr. This estimate is at the low end of the range suggested by Lemeunier *et al.* (1986).

Assuming 10 generations per year for *D. simulans* and *D. yakuba,* and a value of 0.645 Myr for N_e generations, the estimated haploid effective population size of either species is $N_e = 6.5 \times 10^6$ (and hence 3.25×10^6 for the diploid population size). This estimate is in excellent agreement with the value of 2×10^6 suggested for *D. simulans* by Berry, Ajioka, and Kreitman (1991). The value $\gamma = 9.95$ in (2) implies that the average selection coefficient for advantageous or mildly deleterious amino acid replacements in *Adh* is $\sigma = \gamma/N_e = 1.53\times10^{-6}$ per generation. That is, only a very small average selection coefficient is required to account for the observed lack of replacement polymorphisms (or, equivalently, the excess of fixed replacements) in the comparison of *D. simulans* with *D. yakuba*.

Incidentally, the estimate of 1.5×10^{-9} mutations per silent site per generation, derived from Rowan and Hunt (1991) and the assumption of 10 generations per year, compares well with the rule of thumb (Drake 1991) that, in metazoans, the overall mutation rate is roughly one mutation per genome per generation. For the *Drosophila* genome of 165 million base pairs, this implies 6.1×10^{-9} mutations per nucleotide pair per generation. These estimates are quite close, particularly since there might be a slightly smaller substitution rate at silent sites within coding regions than for arbitrary nucleotides.

The analysis of the *Adh* data can also be interpreted in another way. If the rate of replacement mutations is uniform across the coding region of *Adh*, then the overall replacement mutation rate of $\mu_r = 0.013\mu_s$ implies that an average of only about 5.7 codons out of the 257 codon positions in the molecule are susceptible to a favorable amino acid replacement at any one time, with all other replacements at that time being strongly deleterious. The estimate of 5.7 amino acid positions is based on equal mutation rates of each nucleotide. If the mutation rates vary according to nucleotide, the estimated number of amino acids susceptible to favorable replacement at any one time is between 2 and 23 (see discussion below).

The remainder of this paper focuses on the technical details pertaining to the estimates and conclusions summarized above. The main themes are, first, the analysis of joint nucleotide configurations at silent sites from a pair of species in order to estimate the mean mutation rate at silent sites and the species divergence time; and, second, the analysis of the expected probability density of polymorphic site frequencies and of fixed differences at the population level in order to estimate the amount of selection required if all favorable and weakly deleterious replacement mutations have the same selective effect. The population estimates are discussed in the next section, with most of the detail deferred to later in the paper. We then discuss the numerical estimation of γ and μ_r, carry out the analysis of joint nucleotide configurations, and finally give some supplementary comments on certain criticisms of the McDonald-Kreitman test.

Mutational Flux and Fixed Differences

Suppose that new mutations arise with probability $v_N > 0$ per generation in a population of haploid size N. Let $X_{i,k}^N$ be the frequency of the descendants of the i^{th} new mutant allele in the population, where $k = 0, 1, 2, \ldots$ is the number of generations that have elapsed since the original mutation occurred. We assume that the processes $\{X_{i,k}^N : i = 0, 1, \ldots\}$ are non-interacting, as would be the case if the mutations occurred at distinct sites that remain in linkage equilibrium, or if the mutations were sufficiently well spaced in time. The mutations could be selectively advantageous, disadvantageous, or neutral, but in any event we assume that the site frequency processes $\{X_{i,k}^N\}$ are stochastically identical Markov chains. (That is, they have the same transition matrices.) Note that $X_{i,0}^N = 1/N$ for each i, since each new process begins with a single mutation. The states 0 and 1 are absorbing states that represent, respectively, the loss of the new allele and its fixation.

Define $T_{i,a}^N = \min\{k : X_{i,k}^N = a\}$ as the number of generations until the i^{th} process attains the frequency a for the first time, and set $T_{i,a}^N = \infty$ if the allele is fixed or lost before this occurs. For the sake of brevity, let $X_k^N = X_{i,k}^N$ and $T_a^N = T_{i,a}^N$ refer to a typical process $X_{i,k}^N$. Then $P(T_1^N < T_0^N)$ is the probability that a new mutant allele is fixed in the population before it is lost.

We apply a diffusion approximation for the discrete processes $\{X_{i,k}^N\}$. The diffusion process is denoted $\{X_t\}$, where time is scaled in units of N generations (i.e., $t = k/N$ for large N), and the infinitesimal generator of $\{X_t\}$ is assumed to be of the form

$$L_x = \frac{1}{2}b(x)\frac{d^2}{dx^2} + c(x)\frac{d}{dx} \tag{3}$$

where $b(x)$ and $c(x)$ are continuous functions on $[0, 1]$. The operator L_x in (3) can be written in the form

$$L_x = \frac{d}{dm(x)}\frac{d}{ds(x)} \tag{4}$$

by introducing an integrating factor. The functions $dm(x)$ and $s(x)$ are called the *speed measure* and the *scale function* of L_x, respectively (Ewens 1979). Later in the paper we derive a diffusion approximation for the expected number of processes $\{X_{i,k}^N\}$ at equilibrium whose allele frequencies are in the range $(p, p+dp)$, and we also calculate the equilibrium rate at which mutant alleles become fixed.

Now, assume that the processes $\{X_{i,k}^N\}$ correspond to two-allele haploid Wright-Fisher models (without mutation) in which organisms carrying a new mutant allele have fitness $w_N = 1 + \sigma_N$ relative to those carrying the non-mutant allele. If $N\sigma_N \to \gamma$ as $N \to \infty$, then the conditions for a diffusion approximation hold with

$$b(x) = \tfrac{1}{2}x(1-x) \quad \text{and} \quad c(x) = \gamma x(1-x) \tag{5}$$

(Ewens 1979). The scale and speed measure (4) are then

$$s(x) = \frac{1 - e^{-2\gamma x}}{2\gamma} \quad \text{and} \quad dm(x) = \frac{2e^{2\gamma x}\,dx}{x(1-x)} \tag{6}$$

where $s(x)$ is normalized so that $s'(0) = 1$.

In general, the hitting times $T_a = \min\{t : X_t = a\}$ for a diffusion process X_t and the scale function $s(x)$ in (4) are related by the identity

$$P(T_a < T_0 \mid X_0 = x) \;=\; \frac{s(x) - s(0)}{s(a) - s(0)}, \quad 0 < x < a$$

If $\{X_{i,k}^N\}$ are the two-allele haploid Wright-Fisher models of (5–6), then as $N \to \infty$

$$P(T_1^N < T_0^N) \sim P(T_1 < T_0 \mid X_0 = \frac{1}{N}) \tag{7a}$$

$$\sim \frac{s(1/N) - s(0)}{s(1) - s(0)} \sim \frac{1}{N}\frac{2\gamma}{1 - e^{-2\gamma}} \tag{7b}$$

(Moran 1959; recall that $X_0^N = 1/N$). Note that (7a) does not follow from standard diffusion approximation theory, which implies only that the difference between the two probabilities in (7a) converges to zero. In this case, this is trivial because the two terms in (7a) converge to zero individually. Whether (7a) holds for general diploid or dioecious two-allele selection-and-drift models, even those with the same diffusion approximation (5), is apparently still an open problem (Moran 1959; T. Nagylaki, personal communication).

Next, assume

$$v_N \to \mu \qquad \text{as } N \to \infty \tag{8}$$

so that μ is the limiting mutation rate per generation at which new mutant alleles arise. For arbitrary processes $\{X_{i,k}^N\}$ satisfying (7a) and $s'(0) = 1$, we prove later in this paper that the limiting density of polymorphic mutant alleles with population frequencies in the range $(x, x + dx)$ is

$$\mu \, \frac{s(1) - s(x)}{s(1) - s(0)} \, dm(x) \tag{9}$$

for $s(x)$ and $dm(x)$ in (4). This means that the expected number of mutant alleles with population frequency p in the range $0 < p_1 < p < p_2 < 1$ is the integral of (9) over the interval (p_1, p_2). (More precisely, the limiting distribution of the population frequencies of surviving non-fixed mutant alleles is a Poisson random field with (9) as its mean density.) We also show that the limiting flux of the processes $\{X_{i,k}^N\}$ into the state 1 (i.e., into fixation) is given by

$$\frac{\mu}{s(1) - s(0)} \tag{10}$$

in the time scale of the diffusion. That is, (10) is the limiting number of mutant alleles per N generations that become fixed. Note that v_N and μ in (8) are expressed as rates per generation, whereas (10) is the rate per N generations. The difference in time scale reflects the fact that most of the new mutant allele processes become lost in the first few generations.

Frequencies of Wright-Fisher alleles. For the two-allele Wright-Fisher model (5–6) with $\gamma \neq 0$, the equilibrium flux of fixations (10) and the limiting density (9) of non-fixed mutant allele frequencies take the respective forms

$$\mu \, \frac{2\gamma}{1 - e^{-2\gamma}} \quad \text{and} \quad 2\mu \, \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \, \frac{dx}{x(1-x)} \tag{11}$$

where $\gamma > 0$ indicates that the mutant alleles are favorable and $\gamma < 0$ that they are unfavorable. In the neutral case ($\gamma = 0$), the limiting fixational flux and polymorphic frequency density in (11) are

$$\mu \quad \text{and} \quad 2\mu \, \frac{dx}{x} \tag{12}$$

respectively. The expressions in (11) and (12) are displayed in Table 1. Finally, from (12), the expected number of neutral alleles with frequencies p in the range $p_1 < p < p_2$ is given by

$$2\mu \, \log(p_2/p_1)$$

This expression differs from the corresponding expected value in the infinite alleles model (Ewens 1972, 1979). However, the Poisson random field model of the preceding section is not the same as the infinite alleles model. The alleles in (12) do not compete with one another, and there are no constraints on the sum of the frequencies $\{X_{i,k}^N\}$. The processes $\{X_{i,k}^N\}$ are also unaffected by subsequent mutations. This model also differs from the infinite sites model of Watterson (1975), since the processes $\{X_{i,k}^N\}$ are assumed to be independent (i.e., in linkage equilibrium). A model of unlinked or independent sites might be preferable on general grounds to the tightly-linked sites of the infinite sites model, not only because of chromosomal recombination, but also because of the likely frequent occurrence of short-segment gene conversion events in both prokaryotes and eukaryotes (Sawyer 1989, Smith *et al.* 1991, Hilliker *et al.* 1991).

For more general processes $\{X_{i,k}^N\}$, where the connection condition (7a) may not hold, the relations (9) and (10) are still valid provided that v_N in (8) is divided by the ratio of the two probabilities in (7a). Sewall Wright (1938) derived a distribution similar to (9) as an approximation to a quasi-stable distribution for one allele under selection and irreversible mutation. Wright's (1938) problem involved a transient distribution for a single allele, and does not carry over to the equilibrium distribution (9) for a random field of alleles.

Between-species comparisons. In order to analyze differences between species, we assume that a single population had become separated into two disjoint and reproductively isolated subpopulations or species at a time t_{div} in the past, measured in terms of the diffusion time scale (i.e., the split occurred $t_{div} \times N$ generations before the present). Both subpopulations are assumed to have haploid effective size N. If mutations that occur after the population split can be distinguished, and if the number of fixations that have occurred in the two species can be approximated by $2t_{div}$ times the equilibrium flux of fixations, then the number of mutations that correspond to fixed differences between the two species is $2t_{div}$ times the first expressions in (11) and (12). An important quantity is the ratio of the expected number of polymorphic alleles in the frequency range $(x, x + dx)$ to the expected number

Table 1. Population fixational flux and polymorphic densities.

	Equilibrium flux of fixations	Limiting density of freqs. of mutant nucleotides
Neutral	μ_s	$2\mu_s \dfrac{dx}{x}$
Favored ($\gamma > 0$) Unfavored ($\gamma < 0$)	$\mu_r \dfrac{2\gamma}{1-e^{-2\gamma}}$	$2\mu_r \dfrac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}} \dfrac{dx}{x(1-x)}$

of fixed differences, which equals

$$\frac{\big(s(1)-s(x)\big)\,dm(x)}{2t_{div}} \;=\; \frac{1-e^{-2\gamma(1-x)}}{t_{div}\,2\gamma(1-x)}\,\frac{dx}{x}$$

In the next section we will use this expression as a basis for estimating the average value of γ for replacement differences between species.

Sampling Formulas and Parameter Estimates

Suppose that two species diverged $t_{div}N_e$ generations ago, and that both have the same haploid effective population size N_e. Assume that the mutation rate for silent sites in the coding region of a particular gene is μ_s per gene per generation, and that the mutation rate for nonlethal replacement mutations is μ_r per gene per generation. Assume further that (i) all new replacement mutations bestow equal fitness $w = 1+\gamma/N_e$, (ii) each new mutation since the divergence of the species occurred at a different site (in particular, the gene has not been saturated with mutations), and (iii) different sites remain in linkage equilibrium. Under these assumptions, the fixational flux and the expected frequency densities of mutant nucleotides at silent and replacement sites in a single random-mating population are those given in Table 1.

Assume further that the number of polymorphic sites at the present time that are destined to become fixed, and the number of site polymorphisms surviving from the time of speciation, can be neglected in comparison with the number of fixed differences between the species and the number of sites that are presently polymorphic. Then the expected numbers of fixed differences between the two species at the present time are

$$2\mu_s t_{div} \quad \text{and} \quad 2\mu_r t_{div}\frac{2\gamma}{1-e^{-2\gamma}} \tag{13}$$

for silent and replacement sites respectively. Note that the second expression in (13) is much more sensitive to γ if $\gamma < 0$ than if $\gamma > 0$. For example, the factor multiplying $2\mu_r t_{div}$ is 4.1×10^{-8} if $\gamma = -10$, but only 20 if $\gamma = 10$. Thus, if the two expressions in (13) are of comparable size, then γ cannot be strongly negative. Note that $\mu_r \le 10\mu_s$ even if only 10% of the sites in a coding region are silent and all amino acid replacements are advantageous or mildly deleterious. If the two expressions in (13) are equal and $\mu_r/\mu_s \le 10$, then $\gamma \ge -1.81$. If γ is large and positive and the two expressions are equal, then $\mu_r/\mu_s \approx 1/(2\gamma)$.

Similarly, the expected mutant frequency densities at silent and replacement polymorphic sites are

$$\begin{aligned} d\nu_s(x) &= 2\mu_s\frac{dx}{x} \quad \text{and} \\ d\nu_r(x) &= 2\mu_r\frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}}\frac{dx}{x(1-x)} \end{aligned} \tag{14}$$

respectively in either population.

Now, suppose that we have aligned DNA sequences from m chromosomes from the first species and n chromosomes from the second species. The next step is to convert the population level estimates (13–14) to sample estimates. Suppose that a mutant nucleotide has population frequency x at a site. Then a random sample of m chromosomes from that population will be monomorphic for the

mutant nucleotide with probability $q_m(x) = x^m$, and will be polymorphic at that site with probability $p_m(x) = 1-x^m-(1-x)^m$. Thus the expected number of silent polymorphic sites in a random sample of size m is

$$\int_0^1 p_m(x)\, d\nu_s(x) \;=\; 2\mu_s \int_0^1 \frac{1-x^m-(1-x)^m}{x}\, dx$$
$$\;=\; 2\mu_s \sum_{k=1}^{m-1} \frac{1}{k} \qquad (15)$$

for $d\nu_s(x)$ in (14). In fact, assuming linkage equilibrium between sites, the number of silent polymorphic sites is a Poisson random variable whose mean (and hence also variance) is given by (15) (see below). The expression (15) is the same as WATTERSON's (1975) formula for the expected number of polymorphic sites in a sample of size m from the infinite sites model. However, the variance of the number of silent polymorphic sites in the infinite sites model is larger than the variance in the Poisson model (15).

The number of silent polymorphic sites in both samples together is then

$$2\mu_s\big(L(m)+L(n)\big) \qquad \text{where}$$
$$L(m) \;=\; \sum_{k=1}^{m-1} \frac{1}{k} \;\approx\; \log m \qquad (16)$$

The estimate (16) for the number of silent polymorphic sites agrees with the observed numbers to within 7% for each of the three pairwise species comparisons using the *Adh* data of MCDONALD and KREITMAN (1991a), assuming the values of μ_s estimated from the joint configurations in the next section.

Similarly, a silent site will be monomorphic in a sample of size m if it is fixed in the population, but may also be monomorphic in a sample by chance if it is polymorphic in the population. Under the assumptions of (13), the number of silent sites in a sample of size m that are monomorphic for a mutant nucleotide is Poisson with mean

$$\mu_s t_{div} + \int_0^1 x^m\, 2\mu_s \frac{dx}{x} \;=\; \mu_s\left(t_{div} + \frac{2}{m}\right)$$

and the number of silent fixed differences between the two samples is

$$2\mu_s\left(t_{div} + \frac{1}{m} + \frac{1}{n}\right) \qquad (17)$$

However, the estimate (17) is about 45% too large for *D. yakuba* versus *D. simulans* and *D. yakuba* versus *D. melanogaster*, and it is more than 5 times too large for *D. melanogaster* versus *D. simulans* (Table 9). The most likely reason for these discrepancies is that the silent polymorphic sites in the present populations that are destined to become fixed are counted in $2\mu_s t_{div}$ in (17) even though they are not yet fixed. The discrepancy is not large except for the comparison *D. melanogaster* versus *D. simulans*, and it only affects the estimate of μ_r in (1–2). A correction for this overestimation is included in the calculation of μ_r in (2).

By a similar argument, the number of replacement sites that are monomorphic for a mutant nucleotide is Poisson with mean

$$\mu_r \frac{2\gamma}{1-e^{-2\gamma}}\, t_{div} + \int_0^1 x^m\, d\nu_r(x)$$

for $d\nu_r(x)$ in (14). The number of replacement fixed differences between the two samples is then Poisson with mean

$$2\mu_r \frac{2\gamma}{1-e^{-2\gamma}}\big(t_{div} + G(m) + G(n)\big) \qquad (18)$$

where

$$G(m) \;=\; \int_0^1 x^{m-1} \frac{1-e^{-2\gamma(1-x)}}{2\gamma(1-x)}\, dx$$

Note $G(m) \le 1/m$ for $\gamma > 0$. By the same reasoning, the expected number of polymorphic replacement sites in a sample is

$$2\mu_r\big(H(m)+H(n)\big) \qquad (19)$$

where

$$H(m) \;=$$
$$\int_0^1 \frac{1-x^m-(1-x)^m}{x(1-x)} \frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}}\, dx$$

The expression $H(m)$ varies by only a factor of two for $\gamma > 0$, since the inequality $A \le (1-e^{-\gamma A})/(1-e^{-\gamma}) \le 1$ for $0 < A < 1$ implies

$$L(m) \;\le\; H(m) \;\le\; 2L(m), \qquad \gamma > 0$$

for $L(m) = \sum_{k=1}^{m-1} 1/k$. However, $H(m)$ will be smaller if γ is large and negative.

Table 2. Sampling formulas[a]

	Fixed differences	Polymorphic sites
Neutral	$2\mu_s\left(t_{div}+\frac{1}{m}+\frac{1}{n}\right)$	$2\mu_s\big(L(m)+L(n)\big)$
Selected	$2\mu_r\frac{2\gamma}{1-e^{-2\gamma}}\big(t_{div}+G(m)+G(n)\big)$	$2\mu_r\frac{2\gamma}{1-e^{-2\gamma}}\big(F(m)+F(n)\big)$

where $L(m)=\sum_{k=1}^{m-1}\frac{1}{k}$ and

$$F(m) = \int_0^1 \frac{1-x^m-(1-x)^m}{1-x}\frac{1-e^{-2\gamma x}}{2\gamma x}\,dx,$$

$$G(m) = \int_0^1 (1-x)^{m-1}\frac{1-e^{-2\gamma x}}{2\gamma x}\,dx$$

[a] Expected numbers for samples of m genes from one species and n genes from a second species.

The ratio of the number of replacement polymorphic sites to the number of replacement fixed differences is given by the ratio of (19) to (18), which is

$$\frac{F(m)+F(n)}{t_{div}+G(m)+G(n)} \tag{20}$$

where

$$F(m) = \frac{1-e^{-2\gamma}}{2\gamma}H(m)$$
$$= \int_0^1 \frac{1-x^m-(1-x)^m}{1-x}\frac{1-e^{-2\gamma x}}{2\gamma x}\,dx$$

after a change of variables in x. The four basic sampling formulas are summarized in Table 2.

Note that $F(m)\sim L(m)/\gamma$ and $G(m)\to 0$ as $\gamma\to\infty$ for $L(m)=\sum_{k=1}^{m-1}1/k$ in (16). Thus the expression in (20) is asymptotic to

$$\frac{L(m)+L(n)}{t_{div}\gamma} \qquad \text{as } \gamma\to+\infty \tag{21}$$

Estimating parameters. We estimate γ by setting the expression (20)

$$\frac{F(m)+F(n)}{t_{div}+G(m)+G(n)} \tag{22}$$
$$= \frac{\text{Numb. observed repl. polymorphisms}}{\text{Numb. observed repl. fixed differences}}$$

for the ratio of the numbers of the observed replacement polymorphic sites and replacement fixed differences. (Note that μ_r cancels in this particular ratio. If $\gamma>0$ is sufficiently large, the approximation (21) can be used instead.) If the number of replacement polymorphic sites is zero (as it is for the *D. simulans* versus *D. yakuba* comparison), then we use 1/2 as a conservative value to replace the zero. For example, if there are no replacement polymorphic sites and 6 fixed replacement differences (as is the case for *D. simulans* versus *D. yakuba*), we set the expression in (20) equal to $\frac{1}{2}/6=1/12$ and solve for γ. Using the estimate $t_{div}=5.8$ derived in the next section, the solution is $\gamma=9.95$. (The approximation (21) gives $\gamma=11.0$ in this case.) Since there are no replacement polymorphic sites for *D. simulans* versus *D. yakuba*, this value is a lower bound for the average value of γ. Any larger value would also be consistent with the data.

The replacement mutation rate μ_r can be estimated by setting the ratio of the expected numbers of replacement and silent fixed differences in (18)

and (17)

$$\frac{\mu_r\, 2\gamma}{\mu_s(1-e^{-2\gamma})}\frac{t_{div}+G(m)+G(n)}{t_{div}+1/m+1/n} \quad (23)$$
$$= \frac{\text{Numb. observed repl. fixed differences}}{\text{Numb. observed silent fixed differences}}$$

for the ratio of the numbers of the observed replacement and silent fixed differences. Setting the theoretical ratio (23) equal to the observed ratio 6/17 for *D. simulans* and *D. yakuba* yields $\mu_r = 0.037 = 0.018\mu_s$. (Recall that only regular silent sites are counted in the estimation of μ_s; see the next section.) However, as noted earlier, both $2\mu_s t_{div}$ and (17) overestimate the observed number of silent fixed differences, as well as the estimated number of silent fixed differences using the probability distribution of joint configurations (see the next section). In both cases, the estimate (17) is too large by a factor of about 1.4 (Table 9). Since replacement polymorphisms with favorable mutant nucleotides are likely to become fixed faster than silent polymorphisms, we apply a correction of 1.4 to μ_s but not to μ_r in (23). This correction leads to the estimate $\mu_r = 0.0265 = 0.013\mu_s$ for *D. simulans* versus *D. yakuba*. The interpretation of μ_r in terms of an average of 5.7 amino acids susceptible to favorable replacement is discussed at the end of the next section.

Under our assumptions, the numbers of silent and replacement fixed differences and polymorphic sites are independent Poisson random variables (see below). Maximum likelihood estimators for μ_r and γ (along with associated confidence intervals) can be found by setting the expressions (18) and (19) (respectively) equal to the observed numbers of replacement fixed differences and replacement polymorphic sites. The maximum likelihood estimator for γ is the same as the estimator (22) for γ. However, we prefer to estimate μ_r in terms of μ_s and the more numerous silent site data rather than use the maximum likelihood estimator for μ_r in this case.

Divergence Times and Mutation Rates

We treat a K-fold degenerate silent site as a neutral Wright-Fisher model with K types, no selection, and mutation rate u_{ij} per generation from type i to type j. If $u_{ij} = u_j$ depends only on j, and if $\nu_{ij} = N_e u_{ij} = \nu_j = N_e u_j$ is the scaled mutation rate where N_e is the effective haploid population size, then Wright's (1949) formula states that the equilibrium population frequencies $p_1, \ldots, p_K$ of the K types are random with probability density

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} p_1^{\alpha_1-1} p_2^{\alpha_2-1}\ldots p_K^{\alpha_K-1} \quad (24)$$

for large N_e. In (24), $\alpha_j = 2\nu_j = 2N_e u_j$, $\alpha = \sum_{i=1}^{K}\alpha_i$, and $\Gamma(\alpha)$ is the gamma function. It follows from (24) that $E(p_i) = \alpha_i/\alpha$ for all i. If $\alpha_1 = \alpha_2 = \cdots = \alpha_K$, then $\alpha = K\alpha_i$ and $E(p_i) = 1/K$ for $1 \le i \le K$.

However, nucleotide frequencies at third-position sites in fourfold degenerate codons tend to be far from uniform (see Table 3).

Table 3. Nucleotide frequencies at 4-fold degenerate sites in the *Adh* region[a]

Species	Seqs	Sites	T	C	A	G
D. simulans	6	109	0.18	0.62	0.06	0.14
D. melanogaster	12	108	0.17	0.60	0.07	0.16
D. yakuba	12	110	0.14	0.61	0.07	0.18

[a] Data from McDonald and Kreitman (1991a)

Assuming that the sites are independent, departures from an even distribution are highly significant in all cases in Table 3 ($P < 10^{-11}$, 3 d.f.), although the observed nucleotide frequencies do not differ significantly among the three species ($P = 0.10$, 6 d.f.). The mutation model (24) with variable α_i provides a significantly better fit to the distribution of nucleotides at silent sites in the *Adh* data than does (24) with $\alpha_i = \alpha_1$ ($P < 10^{-15}$, 3 d.f., in all three species, assuming independent sites). Although a mutation model with rates a function of the final nucleotide, rather than the original, is somewhat artificial, it is reasonable in the analysis of these data since only the existing nucleotide sequences are known. Furthermore, since the initial and final nucleotides may very well be correlated (e.g., because of a transition or transversion bias), the model with variable α_i is probably reasonably robust. In any event, it provides a significantly better fit to the *Adh* data.

Assuming Wright's model (24) at K-fold degenerate silent sites, the equilibrium distribution

Table 4. Sample configurations.

Configuration	Probability from (25)
TTCAAG	$\frac{6!}{2!\,1!\,2!\,1!}\,\frac{\alpha_1(\alpha_1+1)\alpha_2\alpha_3(\alpha_3+1)\alpha_4}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)(\alpha+5)}$
AAAGGG	$\frac{6!}{0!\,0!\,3!\,3!}\,\frac{\alpha_3(\alpha_3+1)(\alpha_3+2)\alpha_4(\alpha_4+1)(\alpha_4+2)}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)(\alpha+5)}$
CCCCCA	$\frac{6!}{0!\,5!\,1!\,0!}\,\frac{\alpha_2(\alpha_2+1)\dots(\alpha_2+4)\alpha_3}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)(\alpha+5)}$
TTTTTC CAAAGG	$\frac{12!}{5!\,2!\,3!\,2!}\,\frac{\alpha_1(\alpha_1+1)\dots(\alpha_1+4)\alpha_2(\alpha_2+1)\alpha_3(\alpha_3+1)(\alpha_3+2)\alpha_4(\alpha_4+1)}{\alpha(\alpha+1)(\alpha+2)\dots(\alpha+11)}$
TCCCCC CCCCCC	$\frac{12!}{1!\,11!\,0!\,0!}\,\frac{\alpha_1\alpha_2(\alpha_2+1)\dots(\alpha_2+10)}{\alpha(\alpha+1)(\alpha+2)\dots(\alpha+11)}$

of population nucleotide frequencies is a K-type Dirichlet distribution (24) with parameters $\alpha_i = 2N_e u_i$, where N_e is the haploid effective population size. If a sample of size m is randomly chosen from this model, the probability that the sample will contain m_i nucleotides of type i for $1 \le i \le K$ is then

$$\frac{m!}{m_1!\,m_2!\dots m_K!}\,\frac{\prod_{i=1}^{K}\alpha_i^{(m_i)}}{\alpha^{(m)}}, \quad \alpha = \alpha_1 + \cdots + \alpha_K \tag{25}$$

where $\alpha^{(m)} = \alpha(\alpha+1)\dots(\alpha+m-1)$. Examples of (25) for some particular configurations are given in Table 4.

There are two types of twofold degenerate sites. At sites of the first type, the nucleotide can be either of the two pyrimidines T or C, for which (25) holds with $K = 2$. At sites of the second type, the nucleotide can be either of the two purines A or G, and (25) holds with $K = 2$ and α_1, α_2 replaced by α_3, α_4. The only 3-fold degenerate amino acid, isoleucine, has third-position synonymous nucleotides T, C, and A. Since the codon ATA is almost entirely absent in the *Adh* sequences, we treat 3-fold degenerate sites as 2-fold degenerate in the analysis of silent site distributions, and ignore silent codon positions containing an ATA. Synonymous sites within leucine and arginine codons are of variable degeneracy, since silent mutations in the first (or third) codon position can change the degeneracy in the third (or first) position. Codons for serine fall into two nonoverlapping classes, one 4-fold degenerate and one 2-fold degenerate; we treat these two classes as separate amino acids. We define a *regular silent site* as either any 2-fold or 4-fold degenerate third-position site in an amino acid monomorphic codon position that does not code for leucine or arginine, or else as a third-position site in an amino acid monomorphic isoleucine codon position that does not contain an ATA codon (Sawyer, Dykhuizen, and Hartl 1987). Regular isoleucine silent sites are considered 2-fold degenerate. The analysis of nucleotide variation at silent sites is restricted to regular silent sites.

The scaled mutation rate to nucleotides of type j is $\nu_j = \alpha_j/2$. Since $E(p_i) = \alpha_i/\alpha$ by (24), the overall mean mutation rate at 4-fold degenerate silent sites is

$$\mu_4 = \sum_{i=1}^{4} \alpha_i(\alpha - \alpha_i)/2\alpha, \quad \alpha = \alpha_1 + \cdots + \alpha_4 \tag{26}$$

with the rates

$$\begin{aligned} \mu_{TC} &= \alpha_1\alpha_2/(\alpha_1+\alpha_2) \quad \text{and} \\ \mu_{AG} &= \alpha_3\alpha_4/(\alpha_3+\alpha_4) \end{aligned} \tag{27}$$

at 2-fold degenerate sites. The aggregate silent mutation rate per gene at regular 2-fold and 4-fold degenerate sites is then

$$\mu_s = N_{2,TC}\,\mu_{TC} + N_{2,AG}\,\mu_{AG} + N_4\mu_4 \qquad (28)$$

where $N_{2,TC}$ and $N_{2,AG}$ are the numbers of regular 2-fold degenerate silent sites of each type (TC or AG), N_4 is the number of regular 4-fold degenerate silent sites, and 3-fold degenerate sites are counted in $N_{2,TC}$ unless the codon position contains an ATA.

Table 5 gives the within-species maximum likelihood estimates for $\alpha_1, \ldots, \alpha_4$ based on the configuration probabilities (25) at regular silent sites with all four α_i's varied independently in the maximization. The last two columns in Table 5 give the resulting estimates for μ_4 in (26) and μ_s in (27–28). Inferred 95% confidence intervals are generally on the order of plus or minus half the size of the estimated MLE's. Table 6 gives one-dimensional maximum likelihood estimates for the single-α model $\alpha_i \equiv \alpha_1$. While the values for μ_4 and μ_s in Table 6 are essentially the same as in the four-α model, the model with variable α_i provide a significantly better fit to the data in all three species.

An estimate of divergence time. A time-dependent version of Wright's formula (24) was first derived by Griffiths (1979). Assume that two equilibrium Wright-Fisher K-type populations diverged $t_{div} \times N_e$ generations ago, where N_e is the haploid effective population size. The ancestral population and both offshoot populations are assumed to have the same haploid effective population size and the same mutation structure $u_{ij} \equiv u_j$. Given present nucleotide frequencies $p_1, \ldots, p_K$ in one of the populations, then the present frequencies $q_1, \ldots, q_K$ in the other population are random with probability density

$$P(t,p,q) \;=\; \sum_{b=0}^{\infty} d_b(2t) \sum_{\sum b_i = b} \frac{b!}{b_1! \ldots b_K!} \prod_{i=1}^{K} p_i^{b_i} \times \frac{\Gamma(b+\alpha)}{\Gamma(b_1+\alpha_1)\ldots\Gamma(b_K+\alpha_K)} q_1^{b_1+\alpha_1-1} \ldots q_K^{b_K+\alpha_K-1} \qquad (29)$$

where $\alpha_i = 2N_e u_i$ and $\alpha = \alpha_1 + \cdots + \alpha_K$ (Griffiths 1979). The coefficients $d_b(2t)$ in (29) satisfy

$$d_b(t) \;=\; \sum_{k=b}^{\infty} \frac{(2k+\alpha-1)(-1)^{k-b}(\alpha+b)^{(k-1)}}{k!} e^{-\lambda_k t} \qquad (30a)$$

if $b \geq 1$ and

$$d_0(t) \;=\; 1 - \sum_{k=1}^{\infty} \frac{(2k+\alpha-1)(-1)^{k-1}\alpha^{(k-1)}}{k!} e^{-\lambda_k t} \qquad (30b)$$

where $\alpha^{(m)} = \alpha(\alpha+1)\ldots(\alpha+m-1)$ as before and $\lambda_k = k(k+\alpha-1)/2$. Since $d_b(t) = O(e^{-b^2 t/3})$ for large b, the series (29) converges rapidly in b unless t is small.

Equation (29) has a simple intuitive explanation (Tavaré 1984). In the limit as $N \to \infty$ and $Nu_j \to \alpha_j/2$, all individuals in an equilibrium Wright-Fisher population are the descendants of $b < \infty$ "founding" ancestors that lived $t \times N_e$ generations earlier. The expression $d_b(t)$ in (30ab) is the probability distribution of b for t in the diffusion time scale, under the assumption that mutations break the line of descent (Tavaré 1984). That is, $d_b(t)$ is the probability that all individuals in the present population either have descended without intervening mutation from one of b founding individuals who existed t diffusion time units ago, or else are descended from a mutant ancestor that arose within the last t time units. If the nucleotide frequencies were $p_1, \ldots, p_K$ in the ancestral population, the probability that b_i of the b founding ancestral individuals were of type i for $1 \leq i \leq K$ is $(b!/b_1! \ldots b_K!) \prod p_i^{b_i}$. The unmutated descendants of these b individuals will have the same states at the present time. Thus the nucleotide frequencies $q_1, \ldots, q_K$ at the present time have a Dirichlet distribution (as in Wright's formula (24)) conditioned on a sample of size b having b_i individuals of type i for $1 \leq i \leq K$. However, a K-type Dirichlet distribution with parameters α_i, conditioned that a sample of size b had b_i individuals of type i, is Dirichlet with parameters $b_i + \alpha_i$. Finally, by time reversibility, the joint distribution of two present day populations, connected through an ancestral population t time units ago, is the same as the joint distribution of one population and an ancestral or descendant population $2t$ time units apart. This completes the proof of Griffith's formula (29).

Suppose that a sample of size m is taken from one species, and of size n from another, closely related, species. It follows from (29) that the joint probability that the first sample has m_i bases of type i $(1 \leq i \leq K)$, and that the second sample

Table 5. MLE's from Wright's Formula (25)[a]

Species	α_1 (T)	α_2 (C)	α_3 (A)	α_4 (G)	μ_4	μ_s
D. simulans	0.0101	0.0326	0.0021	0.0096	0.0156	2.33
D. melanogaster	0.0082	0.0259	0.0020	0.0083	0.0130	1.92
D. yakuba	0.0063	0.0281	0.0025	0.0100	0.0135	1.93

[a] Estimated from regular silent sites (see text)

Table 6. MLE's assuming $\alpha_i \equiv \alpha_1$

Species	Normal-theory 95% CI for α_1	μ_4	μ_s
D. simulans	$(0.0104 \pm 0.0065) = (0.0039, 0.0169)$	0.0155	2.26
D. melanogaster	$(0.0087 \pm 0.0052) = (0.0035, 0.0139)$	0.0130	1.87
D. yakuba	$(0.0085 \pm 0.0051) = (0.0041, 0.0136)$	0.0127	1.86

Table 7. *Drosophila* pairwise comparisons

					Model with $\alpha_i \equiv \alpha_1$		
Species	μ_s	t_{div}	95% CI for t_{div}	t_{gene}	μ_s	t_{div}	t_{gene}
simulans vs. *yakuba*	2.05	5.81	(3.28, 8.33)	6.72	1.98	5.55	6.46
melanogaster vs. *yakuba*	1.85	7.08	(4.06, 10.11)	8.18	1.80	6.54	7.85
melanogaster vs. *simulans*	2.07	1.24	(0.48, 1.99)	2.26	2.01	1.23	2.26

has n_j bases of type j $(1 \le j \le K)$, is

$$C_{mn} \sum_{b=0}^{\infty} d_b(2t) \sum_{\sum b_i = b} \frac{b!}{b_1! \ldots b_K!} \times \frac{\prod_{i=1}^{K} \alpha_i^{(b_i)} \, (\alpha_i + b_i)^{(m_i)} \, (\alpha_i + b_i)^{(n_i)}}{\alpha^{(b)} \, (\alpha + b)^{(m)} \, (\alpha + b)^{(n)}} \tag{31}$$

where $C_{mn} = m!/(m_1! \ldots m_K!) \times n!/(n_1! \ldots n_K!)$ and $\alpha^{(m)} = \alpha(\alpha+1)\ldots(\alpha+m-1)$ as before.

Given an aligned sample of DNA sequences from two species, we estimate the divergence time between the two species as follows. First, all regular silent sites (as defined earlier) within the two species are pooled to find maximum likelihood estimates of $\alpha_1, \ldots, \alpha_4$ using Wright's formula (25) for the configuration probabilities. Using the estimated values for α_i, a maximum likelihood estimate of t_{div} is then found using the likelihoods (31) at regular silent sites in the two species. Since (31) depends on t only through $d_b(2t)$, arrays $(\alpha_i + b_i)^{(m_i)}$ etc. can be computed in advance once the α_i are known, independently of t. The most time-consuming part of the maximization is the computation of the coefficients multiplying $d_b(2t)$ in (31) for $K = 4$ and large b.

Estimates of t_{div} using this method and the data of McDonald and Kreitman (1991a) are given in Table 7. Note that our method gives a direct estimate of *species* divergence times, rather than of *gene* divergence times (i.e., the time since the common ancestor of a set of genes, which may be considerably older than the time since the species diverged). In practice, estimates of the average pairwise gene divergence times t_{gene} (Table 7) were computed as a starting point for the maximization of the likelihood for t_{div}. These estimates of t_{gene} tended to be larger than t_{div}, particularly when t_{div} was small. Table 7 also contains 95% confidence intervals for t_{div} based on the one-dimensional likelihood curvature at t_{div}. However, these confidence intervals may overstate the accuracy of the estimates since the α_i were held constant. The last three columns in Table 7 give estimates for t_{div} based on the single-α model $\alpha_i \equiv \alpha_1$. These were quite similar to the estimates in the preceding four columns based on the four-α model.

Watterson (1985, see also Padmadisastra

1988) developed a maximum likelihood method for estimating the divergence time between two populations based on the infinite alleles model at many unlinked loci. If mutation is sufficiently rare so that recurrent mutation at polymorphic sites can be neglected, then this method should give estimates similar to ours, but it does not allow for the possibility of nucleotide-dependent mutation rates. WATTERSON (1985) also discusses five other methods of estimating species divergence times and compares them by computer simulation.

We also calculated bootstrap bias-corrected estimates and bias-corrected 95% confidence intervals (EFRON 1982) for t_{div} in the four-α model (Table 8). These were based on 1000 nonparametric bootstrap simulations for each species pair with α_i fixed. The bias-corrected estimates are quite similar to the MLE's in Table 7, but the bootstrap confidence intervals, particularly the lower limits, are shifted upwards.

Table 8. Bootstrap bias-corrected estimates

Species	t_{div}	95% CI
simulans vs. *yakuba*	5.89	(3.50, 8.90)
melanogaster vs. *yakuba*	7.20	(4.56, 10.52)
melanogaster vs. *simulans*	1.26	(0.81, 1.92)

Comparisons with data. We now compare the observed numbers of silent polymorphic sites and fixed differences with two sets of theoretical estimates of these numbers (Table 9). The first set of theoretical estimates is as follows. Since whether or not a site is polymorphic or a fixed difference depends on its joint configuration in the two samples, the probability that a silent site is polymorphic or a fixed difference can be computed from the joint configuration probabilities (31) once α_i and t_{div} are known. Estimates of the expected numbers of silent polymorphic sites and fixed differences based on the joint configuration probabilities (31), maximum likelihood estimates of α_i and t_{div} (Table 7), and the numbers $N_{2,TC}$, $N_{2,AG}$, and N_4 of regular silent sites of various degeneracies, are given in Table 9. These estimates fit the observed data very closely, which suggests that the neutral joint configuration model (31) fits the data at synonymous sites. Note that a consistently expanding population or deleterious selection would tend to produce fewer polymorphisms than predicted by neutrality, while a contracting population or balancing selection would be likely to have more polymorphisms.

The second set of theoretical estimates are the Poisson-random-field-based sampling estimates derived earlier. The number of silent fixed differences is estimated as follows. By time reversibility, data from two species at the present time can be viewed as two snapshots of a single species separated in time by $2t_{div}$ time units. Under selective neutrality, the rate μ_s of regular silent mutations per individual is the same as the long-term rate of fixations of regular silent differences at the population level, so that the expected number of silent mutations in the population that will eventually become fixed differences is $2\mu_s t_{div}$. However, as we start with one contemporary species and evolve through the ancestral species to the other contemporary species, $2\mu_s t_{div}$ overestimates the number of silent fixed differences by the number of silent mutations that are destined to become fixed in this process but which have not yet had time to fix. There is no corresponding underestimate in the number of polymorphic sites, since an initial polymorphic site in the first contemporary species cannot become a fixed difference as we proceed backwards in time through the ancestral species and then forwards to the second species. (The site remains polymorphic in the present.) Finally, the estimate $2\mu_s t_{div}$ is augmented as in (17) by an estimate of the number of sites that are fixed in the sample but polymorphic in the population.

The estimate (16) for the number of silent polymorphic sites fits the observed numbers quite well. However, the predicted numbers of fixed differences $2\mu_s t_{div}$ and (17) overstate, by factors of between 1.4 and 6.1, both the observed number of fixed differences and the number of fixed differences predicted from the sample configurations. The upward bias of this estimate probably reflects the fact that not all of the new mutations destined to become fixed have as yet become fixed. Based on this argument, we propose that the total number of silent fixed differences in terms of μ_s is approximately the theoretical expression (17) divided by 1.4 for *D. simulans* vs. *D. yakuba*, and the ratio of the number of replacement fixed differences to silent fixed differences can be estimated by 1.4 times the ratio of (18) to (17). Setting the latter expres-

Table 9. Silent polymorphic sites and fixed differences

Species	μ_s	t_{div}	Obs. silent[a] fixed	Obs. silent[a] poly	Est. config.[b] fixed	Est. config.[b] poly	Est. popn.[c] fixed	Est. popn.[c] poly
sim. vs. *yak.*	2.05	5.81	17	21	16.0	19.9	24.8	21.8
mel. vs. *yak.*	1.85	7.08	18	21	17.5	20.5	26.9	22.4
mel. vs. *sim.*	2.07	1.24	1	21	2.0	19.6	6.1	21.9

[a] Regular silent sites only

[b] From the joint configuration probabilities (31)

[c] From (17) and (16)

sion equal to the observed ratio of replacement and silent fixed differences yields the corrected estimate $\mu_r = 0.0265 = 0.013\mu_s$ of (2).

Selective constraints in *Adh* evolution. The low value of the replacement mutation rate μ_r, relative to the silent mutation rate μ_s, may reflect mutations occurring in only a small number of codons at which a favorable amino acid change is possible, with all other changes being strongly detrimental. The actual mutation rate at this small number of susceptible sites may be the same as at silent sites. In the single-α model $\alpha_i \equiv \alpha_1$ for pooled data from *D. simulans* and *yakuba*, the mutation rate from one nucleotide to another is $\nu = \alpha_1/2 = 0.00465$, which suggests that the average number of codons susceptible to a favorable amino acid change at any one time is $n = \mu_r/\nu = 0.0265/0.00465 = 5.7$ out of the 257 total codon positions available. In the more general mutation model, the mutation rate depends on the nucleotide produced by the mutation, and the estimates $n_i = \mu_r/(\alpha_i/2) = 2\mu_r/\alpha_i$ range from an average of 1.8 susceptible codons (for mutations to C) to 23.1 susceptible codons (for mutations to A).

Mutational Flux and Fixed Differences: Derivations

The purpose of this section is to derive the limiting density of the processes $\{X^N_{i,k}\}$ in the frequency interval $[0,1]$, and the limiting rate at which these processes are fixed at the state 1. For each $N \geq 1$, let $\{X^N_{i,k}\}$ $(i = 1, 2, \ldots)$ be a sequence of Markov chains on $\{0, 1/N, 2/N, \ldots, 1\}$ with the same transition matrix. Assume that one of these processes starts at $X^N_{i,0} = 1/N$ with probability $v_N > 0$ in each time step. The endpoints $0, 1$ are assumed to be absorbing states for the Markov chains. Given the initial states $x_N = j/N \to x$, the processes $\{X^N_k = X^N_{i,k}\}$ are assumed to satisfy the conditions

$$\begin{aligned} \lim_{N\to\infty} NE\big(X^N_{k+1} - X^N_k \;\big|\; X^N_k = x_N\big) &= c(x) \\ \lim_{N\to\infty} NE\big((X^N_{k+1} - X^N_k)^2 \;\big|\; X^N_k = x_N\big) &= b(x) \\ \lim_{N\to\infty} NE\big(|X^N_{k+1} - X^N_k|^{2+\delta} \;\big|\; X^N_k = x_N\big) &= 0 \end{aligned} \tag{32}$$

uniformly for $0 \leq x \leq 1$ for some $\delta > 0$, where $b(x)$ and $c(x)$ are continuous functions on $[0,1]$. Let $dm(x)$ and $s(x)$ be the speed measure and scale function of the diffusion operator

$$L_x = \frac{1}{2}b(x)\frac{d^2}{dx^2} + c(x)\frac{d}{dx}$$

(see equation (4)), and assume that the endpoints $0, 1$ are accessible boundaries for L_x (or for the limiting diffusion process X_t). All of these conditions hold for the two-allele selection-and-drift haploid Wright-Fisher models discussed earlier (Ewens 1979).

The probability v_N that a new process $\{X^N_{i,k}\}$ begins in any time step is assumed to satisfy

$$v_N \;\sim\; \frac{\mu}{N\,P(T^N_{a+} < T^N_0)\big(s(a) - s(0)\big)} \qquad \text{as } N \to \infty \tag{33}$$

where $T^N_{a+} = \min\{k : X^N_k \geq a\}$ and $a > 0$ $(0 < a \leq 1)$ is arbitrary. In fact, condition (33) is independent of a for $0 < a \leq 1$ (see below). If the condition $P(T^N_{a+} < T^N_0) \sim P(T_a < T_0 \mid X_0 = 1/N)$ holds as $N \to \infty$ for $T_t = \min\{t : X_t = a\}$ and $s'(0) = 1$, then (33) is equivalent to $v_N \to \mu$.

At equilibrium for fixed N, the expected number of processes $\{X^N_{i,k}\}$ at the state x for $0 < x < N$

is given by

$$B_N(x) = v_N \sum_{k=0}^{\infty} P(X_{k,k}^N = x) \;=\; v_N \sum_{k=0}^{\infty} P(X_k^N = x) \tag{34}$$

where k in (34) represents a time k generations in the past. If $f(x)$ is a function on $[0,1]$ with $f(0) = f(1) = 0$, then the expected number of processes $\{X_{i,k}^N\}$ in the open interval $(0,1)$, weighted by $f(x)$ where x is the present frequency, is

$$\sum_{j=1}^{N-1} f(j/N) B_N(j/N) \;=\; v_N \sum_{k=0}^{\infty} Q_k^N f(1/N)$$

where $Q_k^N f(x) = E\big(f(X_k^N) \;\big|\; X_0^N = x\big)$. Our first result is the limit theorem

Theorem 1. *Suppose under the above assumptions that $f(x)$ is continuous on $[0,1]$ and that $f(x) = 0$ in the interval $[0,a)$ for some $a > 0$. Then*

$$\begin{aligned} &\lim_{N\to\infty} \sum_{j=1}^{N-1} f(j/N) B_N(j/N) \\ &\quad = \lim_{N\to\infty} v_N \sum_{k=0}^{\infty} Q_k^N f(1/N) \\ &\quad = \int_0^1 \mu \frac{s(1) - s(x)}{s(1) - s(0)} f(x)\, dm(x) \end{aligned} \tag{35}$$

where $s(x)$ is the scale and $dm(x)$ the speed measure of the limiting diffusion.

It follows from (35) that the limiting expected number of processes $\{X_{i,k}^N\}$ with frequencies in the range $p_1 \le p \le p_2$ is the integral of the integrand on the right-hand side of (35) over (p_1, p_2), provided that p_1, p_2 are points of continuity for $dm(x)$. In fact, the limiting distribution of $\{X_{i,k}^N\}$ is a Poisson random field with this integrand as its mean density (see the next section).

Proof of Theorem 1. We first show that condition (33) is independent of a. If $0 < 1/N < a < 1$, then X_k^N must first cross a before getting to 1. This leads to the identity

$$\begin{aligned} &P(T_1^N < T_0^N) \\ &\quad = P(T_{a+}^N < T_0^N)\, E\big(P(T_1^N < T_0^N \mid X_0^N = Y)\big) \end{aligned} \tag{36}$$

for $Y = X_\ell^N$, where $\ell = T_{a+}^N$ is the first time at which the frequency $X_\ell^N \ge a$. For any fixed $y < 1$,

$$\begin{aligned} &\lim_{N\to\infty} P(T_1^N < T_0^N \mid X_0^N = y) \\ &\quad = P(T_1 < T_0 \mid X_0 = y) \;=\; \frac{s(y) - s(0)}{s(1) - s(0)} \end{aligned} \tag{37}$$

(Ewens 1979; Ethier and Kurtz 1986). It follows from (37) and (32) that the "overshoot" in (36) can be neglected as $N \to \infty$; i.e., it is sufficient to take $Y = a$ in (36). Hence by (36) and (37)

$$\begin{aligned} &P(T_{a+}^N < T_0^N)\big(s(a) - s(0)\big) \\ &\quad \sim\; P(T_1^N < T_0^N)\big(s(1) - s(0)\big) \end{aligned}$$

as $N \to \infty$ for $0 < a < 1$, and the condition (33) is independent of $a > 0$. In particular, if

$$\begin{aligned} P(T_1^N < T_0^N) &\sim P(T_1 < T_0 \mid X_0 = 1/N) \\ &= \frac{s(1/N) - s(0)}{s(1) - s(0)} \sim \frac{1}{N} \frac{s'(0)}{s(1) - s(0)} \end{aligned}$$

then (33) implies that $v_N \to \mu / s'(0)$.

Similarly, for $a > 0$ in Theorem 1, before any of the processes $\{X_{i,k}^N\}$ get to $x \ge a$, they must first cross a. Thus the sum in (35) equals

$$\begin{aligned} &v_N\, P(T_{a+}^N < T_0^N)\, E\Big(\sum_{k=0}^{\infty} Q_k^N f(Y)\Big) \\ &\quad \sim\; \frac{\mu}{s(a) - s(0)} \frac{1}{N} \sum_{k=0}^{\infty} Q_k^N f(a) \end{aligned} \tag{38}$$

by (33), where $Y = X_\ell^N$ for $\ell = T_{a+}^N \ge a$ as before, since we can neglect the overshoot in (38) for the same reasons as in (36). Now by (32) and Trotter's Theorem (Ethier and Kurtz 1986)

$$\lim_{N\to\infty} Q_k^N f(x_N) \;=\; Q_t f(x) \;=\; E\big(f(X_t) \;\big|\; X_0 = x\big) \tag{39}$$

uniformly for $k/N \to t$ and $x_N \to x$. Thus, if we can show that the sum in (38) converges uniformly in N for $k/N \ge C$, then as $N \to \infty$

$$\begin{aligned} \frac{1}{N} \sum_{k=0}^{\infty} Q_k^N f(a) \;&\to\; \int_0^{\infty} Q_t f(a)\, dt \\ &= \big(s(a) - s(0)\big) \int_a^1 \frac{s(1) - s(x)}{s(1) - s(0)} f(x)\, dm(x) \end{aligned}$$

(Ewens 1979) since $f(x) \equiv 0$ for $x < a$. Since this completes the proof of (35), it remains only to prove that the sum in (38) converges uniformly in N for $k/N \geq C$.

Since the endpoints $0, 1$ are accessible exit boundaries for L_x or $\{X_t\}$, we can choose $t > 0$ and $\delta > 0$ such that

$$\max_{0 \leq x \leq 1} P(T_0 \wedge T_1 \geq t \mid X_0 = x) \;=\; 1 - \delta \;<\; 1$$

where $X \wedge Y = \min\{X, Y\}$. However, by (39) with $f(x) \equiv 1$,

$$\begin{aligned} \lim_{N \to \infty} & P(T_0^N \wedge T_1^N \geq k \mid X_0^N = x_N) \\ &= P(T_0 \wedge T_1 \geq t \mid X_0 = x) \end{aligned}$$

uniformly for $k/N \to t > 0$ and $x_N \to x$. Thus there exists constants $C > 0$ and $\delta > 0$ such that

$$\begin{aligned} \sup_N \max_{0 \leq j \leq N} & P(T_0^N \wedge T_1^N \geq CN \mid X_0^N = j/N) \\ & \leq 1 - \delta/2 \;<\; 1 \end{aligned} \tag{40}$$

The Markov property implies that if CN in (40) is replaced by mCN, then the right-hand side of (40) can be replaced by $(1 - \delta/2)^m$. Thus there exists some $\nu > 0$ such that if $|f(x)| \leq M$ for $0 \leq x \leq 1$,

$$\begin{aligned} |Q_k^N f(x)| \;&\leq\; M P(T_0^N \wedge T_1^N \geq k \mid X_0^N = x) \\ &\leq\; M \Omega e^{-\nu k/N} \end{aligned}$$

for some constant Ω. This provides the necessary uniformity in (38).

We now compute the limiting flux into the absorbing state 1. In any one time step, the probability that a new process $X_{i,k}^N$ is begun with $X_{i,0}^N = 1/N$ and is then eventually is absorbed at 1 is

$$v_N P(T_1^N < T_0^N) \;\sim\; \frac{1}{N} \frac{\mu}{s(1) - s(0)}$$

by (33) with $a = 1$. Since one time unit in the diffusion time scale equals N discrete time steps, we have shown

Theorem 2. *Under the above conditions, the limiting expected number of processes $\{X_{i,k}^N\}$ that are absorbed at 1 in one time unit in the diffusion time scale is*

$$\frac{\mu}{s(1) - s(0)}$$

Poisson Random Fields and a Sampling Theory for Independent Sites

Since the $\{X_{i,k}^N\}$ are independent Markov processes that arrive in a limiting Poisson stream, the limiting distribution of the frequencies $\{X_{i,k}^N\}$ form a Poisson random field by classical arguments (Karlin and McGregor 1966, Sawyer 1976, Karlin and Taylor 1981). In particular, the limiting distribution of the numbers of frequencies $\{X_{i,k}^N\}$ in any given set, as well as the number of processes that have been fixed at 1 by any given time, are Poisson random variables that are independent for nonoverlapping sets.

Given a sample of size m, the population frequencies $\{X_{i,k}^N\}$ of mutant nucleotides at those sites that are polymorphic in the sample form a "randomly censored" version of the original Poisson random field, where a process is "censored" if its site is monomorphic in the sample. If the censoring mechanism is independent for different sites (which follows in this case from linkage equilibrium), then the censored random field is also a Poisson random field. In particular, given a sample of size m, the numbers of silent and replacement fixed and polymorphic sites all have Poisson distributions.

Similar arguments show that, if one has two or more random censoring mechanisms that are independent for different sites but mutually exclusive (i.e., a site cannot survive censoring by more than one mechanism), then the respective censored random fields are independent Poisson random fields. As one example, given a sample of size m, the number of monomorphic sites in the sample and the number of polymorphic sites in the sample are independent Poisson random variables. In this example, a site is censored by the first mechanism if it is polymorphic in the sample and censored by the second mechanism if it is monomorphic in the sample. (The number of sites that are fixed in the population form a separate independent Poisson class.) It follows from this that, if one has samples of sizes m and n from two populations, then the number of fixed differences between the two samples and the number of sites that are polymorphic in either sample are realizations of independent Poisson random variables.

As a second example, given a sample of size 5 from a single population, the number of sites that

display a "(32)" polymorphism (i.e., with three sequences having one nucleotide and two sequences having a second nucleotide) and the number of sites that display a "(41)" polymorphism (i.e., with four sequences having one nucleotide and one sequence with a second nucleotide) are independent Poisson random variables. Here the censoring mechanisms are to reject a site if it does not have a (32) (respectively (41)) nucleotide polymorphism in the sample, where the mutant nucleotide could be either nucleotide.

Discussion and Additional Comments

The approach to polymorphism and divergence presented in this paper provides a framework for the quantitative analysis and interpretation of DNA sequence variation within and between species. It also provides a theoretical basis for the use of a recently proposed 2×2 contingency table test for DNA sequences that compares polymorphisms and fixed differences with silent sites and amino acid replacements (McDonald and Kreitman 1991a), as well as providing estimates for the relevant parameters. Hypothesis tests for 2×2 contingency table data are well understood and quite robust, and this approach may be the most powerful and generally applicable test yet devised for detecting whether or not selection has been active in the recent evolutionary history of particular proteins.

One the other hand, the sample nucleotide configurations also contain a tremendous amount of information that can be interpreted as in the approach presented here. In particular, the joint sample configurations at silent sites can be used to estimate both the synonymous mutation rate μ_s and the species divergence time t_{div}. Both parameter estimates are scaled in terms of the effective population size N_e, and if independent information is available (as it is in this case from Hawaiian *Drosophila*), then all three parameters can be estimated. Given t_{div}, the 2×2 contingency table provides estimates of μ_r/μ_s (the ratio of aggregate mutations rates for synonymous and replacement sites) and for γ (the average selection coefficient among favored and mildly deleterious replacement mutations). Our analysis is based on the assumption that the nucleotide sites are independent (i.e., in linkage equilibrium). Whether this assumption holds to an acceptable approximation has to be determined by appropriate statistical tests on a case by case basis.

Other approaches to the analysis of DNA sequence variation within and between species were suggested as alternatives to the 2×2 contingency table test by Graur and Li (1991) and Whittam and Nei (1991). These tests appear to be less powerful statistically than the 2×2 contingency table test, and may be subject to objections about their underlying genetic assumptions (McDonald and Kreitman 1991b). Furthermore, the test statistics in the proposed alternatives are assumed to have a sampling distribution that is normal, when in fact the sampling distributions are unknown in most cases. As one example, Graur and Li (1991) compare the observed number ($k = 2$) of replacement polymorphisms within all three species in McDonald and Kreitman's (1991a) data with the value of a test statistic $K = K_1 + K_2 + K_3$ (our notation), where each K_i is the number of segregating sites in an infinite sites model (Watterson 1975). Separate parameter estimates are used within each species, and the random variables K_i are independent. Watterson (1975) gives the mean and variance of K_i in terms of its parameters. Watterson also gives a moment generating function for K_i that can be used to infer its exact distribution (as a sum of independent geometrically distributed random variables with different parameters), from which the exact one-sided P-value $P(K \leq 2)$ can be calculated. For one set of parameter values used by Graur and Li (1991), $K = 10.32 \pm 4.04$ (mean and standard deviation), so that a normal approximation for K leads to a one-sided P-value $P = 0.020$ for the observed $K = 2$, while the exact $P(K \leq 2) = 0.00981$ from the theoretical distribution. In this example the difference in the P-values is only twofold, but it is in the direction of making the Graur and Li statistic less powerful, and the discrepancy might be larger in other cases. Similar problems arise for a test proposed by Whittam and Nei (1991), which is based on a ratio of test statistics each of whose sampling distributions is known only approximately. For ratio statistics, there is no guarantee that approximate P-values will even be conservative. In general, one should be careful about using a normal approximation for P-values unless one is sure that the sampling distribution of the test statistic is at least approximately normally distributed.

We would like to thank Richard Hudson, Tom Nagylaki, Zé Ayala, and two referees for many helpful comments. This work was supported by National Science Foundation Grant DMS-9108262 and National Institutes of Health research grant GM44889 (SAS), and by National Institutes of Health research grants GM30201, GM33741, and GM40322 (DLH).

Literature Cited

Begun, D. J. and C. F. Aquadro, 1991 Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* **129:** 1147–1158.

Berry, A. J., J. W. Ajioka, and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129:** 1111–1117.

Caccone, A., G. D. Amato, and J. R. Powell, 1988 Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. *Genetics* **118:** 671–683.

Drake, J. W., 1991 Spontaneous mutation. *Annu. Rev. Genet.* **25:** 125–146.

Efron, B., 1987 Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82:** 171–200.

Ethier, S. N. and T. G. Kurtz, 1986 *Markov Processes*. Wiley and Sons, New York.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3:** 87–112.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

Graur, D., and W.-H. Li, 1991 Scientific correspondence. *Nature* **354:** 114–115.

Griffiths, R. C., 1979 A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Probab.* **11:** 310–325.

Hartl, D. L., 1989 Evolving theories of enzyme evolution. *Genetics* **122:** 1–6.

Hartl, D. L., and Dykhuizen, D.E., 1981 Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*. *Proc. Nat. Acad. Sci. USA* **78:** 6344–6348.

Hartl, D. L., and S. A. Sawyer, 1991 Inference of selection and recombination from nucleotide sequence data. *J. Evol. Biol.* **4:** 519–532.

Hilliker, A. J., S. H. Clark, and A. Chovnick, 1991 The effect of DNA sequence polymorphisms on intragenic recombination in the *rosy* locus of *Drosophila melanogaster*. *Genetics* **129:** 779–781.

Hudson, R. R., M. Kreitman, and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153–159.

Karlin, S., and J. McGregor, 1966 The number of mutant forms maintained in a population. *Proc. Fifth Berk. Symp. of Math. Stat. Prob.* **4:** 403–414.

Karlin, S., and H. M. Taylor, 1981 *A Second Course in Stochastic Processes*. Academic Press, New York.

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

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, England.

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

Lemeunier, F., J. R. David, L. Tsacas, and M. Ashburner, 1986 The *melanogaster* species group, pp. 147–256 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner and H. L. Carson. Academic Press, New York.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R. C., 1991 Electrophoresis in the development of evolutionary genetics: Milestone or millstone? *Genetics* **128:** 657–662.

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

McDonald, J. H. and M. Kreitman, 1991a Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.

McDonald, J. H. and M. Kreitman, 1991b Scientific correspondence. *Nature* **354:** p116.

Moran, P. A. P., 1959 The survival of a mutant gene under selection. II. *Jour. Australian Math. Soc. I* **1:** 485–491.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. *Nature* **246:** 96–98.

Padmadisastra, S., 1988 Estimating divergence times. *Theor. Popul. Biol.* **34:** 297–319.

Powers, D. A., T. Lauerman, D. Crawford, and L. DiMichele, 1991 Genetic mechanisms for adapting to a changing environment. *Annu. Rev. Genet.* **25:** 629–659.

Rowan, R. G., and J. A. Hunt, 1991 Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for five closely related species of Hawaiian *Drosophila*. *Mol. Biol. Evol.* **8:** 49–70.

Sawyer, S. A., 1976 Branching diffusion processes in population genetics. *Adv. Appl. Probab.* **8:** 659–689.

Sawyer, S. A., 1989 Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6:** 526–538.

Sawyer, S. A., D. E. Dykhuizen, and D. L. Hartl, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Nat. Acad. Sci. USA* **84:** 6225–6228.

Smith, J. M., C. G. Dowson, and B. G. Spratt, 1991 Localized sex in bacteria. *Nature* **349:** 29–31.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26:** 119–164.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7:** 256–276.

Watterson, G. A., 1985 Estimating species divergence times using multi-locus data, pp163–183 in *Population Genetics and Molecular Evolution*, edited by T. Ohta and K. Aoki, Springer-Verlag, Berlin.

Whittam, T. S. and M. Nei, 1991 Scientific correspondence. *Nature* **354:** 115–116.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. *Proc. Nat. Acad. Sci. USA* **24:** 253–259.

Wright, S., 1949 Adaption and selection, pp365–389 in *Genetics, Paleontology, and Evolution*, edited by G. Jepson, G. Simpson, and E. Mayr. Princeton Univ. Press, Princeton, N.J.