J Mol Evol (2003) 57:S154–S164
DOI: 10.1007/s00239-003-0022-3

JOURNAL OF MOLECULAR EVOLUTION

Bayesian Analysis Suggests that Most Amino Acid Replacements in *Drosophila* Are Driven by Positive Selection

Stanley A. Sawyer,[1] Rob J. Kulathinal,[2] Carlos D. Bustamante,[3] Daniel L. Hartl[2]

[1] Department of Mathematics, Washington University, St. Louis, MO 63130, USA
[2] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

Received: 1 August 2002 / Accepted: 12 December 2002

Abstract. One of the principal goals of population genetics is to understand the processes by which genetic variation within species (polymorphism) becomes converted into genetic differences between species (divergence). In this transformation, selective neutrality, near neutrality, and positive selection may each play a role, differing from one gene to the next. Synonymous nucleotide sites are often used as a uniform standard of comparison across genes on the grounds that synonymous sites are subject to relatively weak selective constraints and so may, to a first approximation, be regarded as neutral. Synonymous sites are also interdigitated with nonsynonymous sites and so are affected equally by genomic context and demographic factors. Hence a comparison of levels of polymorphism and divergence between synonymous sites and amino acid replacement sites in a gene is potentially informative about the magnitude of selective forces associated with amino acid replacements. We have analyzed 56 genes in which polymorphism data from *D. simulans* are compared with divergence from a reference strain of *D. melanogaster*. The framework of the analysis is Bayesian and assumes that the distribution of selective effects (Malthusian fitnesses) is Gaussian with a mean that differs for each gene. In such a model, the average scaled selection intensity ($\gamma = N_e s$) of amino acid replacements eligible to become polymorphic or fixed is -7.31, and the standard deviation of selective effects within each locus is 6.79 (assuming homoscedasticity across loci). For newly arising mutations of this type that occur in autosomal or X-linked genes, the average proportion of beneficial mutations is 19.7%. Among the amino acid polymorphisms in the sample, the expected average proportion of beneficial mutations is 47.7%, and among amino acid replacements that become fixed the average proportion of beneficial mutations is 94.3%. The average scaled selection intensity of fixed mutations is $+5.1$. The presence of positive selection is pervasive with the single exception of *kl-5*, a Y-linked fertility gene. We find no evidence that a significant fraction of fixed amino acid replacements is neutral or nearly neutral or that positive selection drives amino acid replacements at only a subset of the loci. These results are model dependent and we discuss possible modifications of the model that might allow more neutral and nearly neutral amino acid replacements to be fixed.

Key words: Polymorphism/divergence — Selective neutrality — Positive selection — Beneficial/deleterious mutations — Poisson random field — Markov chain Monte Carlo (MCMC)

Introduction

Modern population genetics is faced with the challenge of interpreting ever-increasing amounts of DNA sequence data. The objective is to understand

Correspondence to: Daniel L. Hartl; *email:* dhartl@oeb.harvard.edu

both pattern and process from DNA sequences. By *pattern* we mean the phylogenetic relationships among different species of organisms, while by *process* we mean the interplay among various evolutionary forces in transforming genetic polymorphisms within species into genetic divergence between species.

One obstacle to understanding evolutionary process is that the major evolutionary forces of mutation, migration, selection, and random genetic drift are confounded. The result is that, in interpreting sequence data, it is difficult to specify any particular force, or combination of forces, that could account for the data to the exclusion of all other possibilities. Another obstacle to rigorous inference from sequence data is the usually unknown effect of demographic factors such as population subdivision, changes in population size, or founder effects. A third difficulty is in the relative magnitude of the evolutionary forces. Almost every persistent evolutionary force depends on the product of a relatively large number (the effective population size, N_e), whose magnitude is usually unknown, and a relatively small number (mutation, selection, recombination, gene conversion), whose magnitude is usually also unknown (Lewontin 1974).

In spite of these obstacles, a great deal of progress has been made in interpreting DNA sequences. One fruitful approach has been deducing the implications of the neutral theory of molecular evolution (Kimura 1983) with respect to polymorphism and divergence and developing traditional frequentist statistics to test for departures from the neutral expectations. The conceptual breakthrough that opened this approach is due to Ewens (1972), who deduced the expected frequencies of multiple neutral alleles in a sample of organisms from a natural population. Although the Ewens sampling formula proved to be unsuitable for analyzing the electrophoretic protein polymorphisms for which it was originally intended, its applicability to DNA sequences nevertheless helped stimulate the shift in population genetics theory from the elaboration of the theoretical consequences of somewhat arbitrary models to the development of a theory of inference from observed data.

For tests of neutrality, Tajima's (1989) *D* statistic is among the first and most widely used. It tests the standardized difference between estimates of the scaled mutation rate ($4N_e\mu$) based on the observed number of synonymous nucleotide polymorphisms in a sample of sequences and the observed number of pairwise differences at synonymous sites between sequences in a sample. Other types of tests use coalescent simulations to determine the probabilities of the observed number of haplotypes, haplotype diversity, and other statistics (Hudson 1990; Fu and Li 1993; Hudson et al. 1994; Fay and Wu 2000). Maximum likelihood methods have also been used in the analysis of DNA sequence data, especially using nested models to identify particular genes or regions of genes that depart significantly from neutral expectation (Nielsen and Yang 1998; Yang 1998; Bustamante et al. 2002a). Various *ad hoc* approaches to the analysis of DNA sequences have also proven useful (Templeton 1998, 2002).

Recently, we have employed a Bayesian analysis of polymorphism and divergence in order to infer the relative magnitude of the evolutionary forces that promote amino acid replacements among orthologous proteins in closely related species (Bustamante et al. 2002b). For each gene in the analysis, the data consist of a tabulation of numbers of synonymous nucleotide sites that are polymorphic in one or both species (polymorphism) or different between the species (divergence) as contrasted with numbers of nonsynonymous nucleotide sites that show polymorphism or divergence. This contrast was first used as a statistical test of neutrality by McDonald and Kreitman (1991), who realized that, when the data are arrayed in the form of a 2×2 table, a test for independence evaluates the null hypothesis of selective neutrality. More precisely, it tests whether the evolutionary forces impinging on polymorphism and divergence of synonymous nucleotide sites are the same as those impinging on nonsynonymous nucleotide sites.

Two practical problems with this approach are that samples for single genes are frequently relatively small, which compromises the power of any test for independence, and sometimes they contain one or more cells with a count of zero. Across many genes, however, the tables encompass a great deal of evolutionary information whose interpretation is not dependent on the population frequencies of the polymorphisms. We have taken to calling such 2×2 tables DPRS tables to encourage a standard layout. DPRS is an acronym for the column and row headings of a polymorphism–divergence table, read in clockwise order; that is, from left to right the columns are divergence (D) and polymorphism (P), and from bottom to top the rows are replacement (R; nonsynonymous) and synonymous (S) nucleotides. Heuristic analyses of DPRS tables of *Drosophila* genes have led to the inference that many amino acid replacements are driven by positive selection (Fay et al. 2002; Smith and Eyre-Walker 2002). In this paper, we use a hierarchical Bayesian method to analyze DPRS tables in order to infer the mean and variance of the selection coefficients of amino acid replacements across a set of 56 loci from *D. simulans* and *D. melanogaster*. We also find evidence that a high proportion of amino acid replacements is driven by positive selection.

Theoretical Expectations of Polymorphism and Divergence

Similar to traditional convention, we define K_s as the number of synonymous nucleotide substitutions that are fixed differences between a pair of species and S_s as the number of synonymous nucleotide substitutions that are polymorphic in the species. We also define K_a and S_a as the corresponding numbers of nonsynonymous substitutions (amino acid replacements). In the standard layout of the DPRS table, the entries in the top row are K_s and S_s, and in the bottom row they are K_a and S_a.

The expected values for the entries in a DPRS table can be deduced from the equilibrium flux of fixations and the limiting probability density of polymorphic nucleotide substitutions (Sawyer and Hartl 1992). Assuming independence between nucleotide sites evolving under mutation, selection, and random genetic drift, the random variables K_s, S_s, K_a, and S_a are distributed as independent Poisson distributions with means given by

$$E(K_s) = \theta_s\left(t + \frac{1}{m} + \frac{1}{n}\right) \quad (1)$$

$$E(S_s) = \theta_s[L(m) + L(n)] \quad (2)$$

$$E(K_a) = \theta_a\left(\frac{2\gamma}{1 - e^{-2\gamma}}\right)[t + G(m) + G(n)] \quad (3)$$

$$E(S_a) = \theta_a\left(\frac{2\gamma}{1 - e^{-2\gamma}}\right)[F(m) + F(n)] \quad (4)$$

where

$$L(n) = \sum_{i=1}^{n-1} \frac{1}{i}$$

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

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

In these expressions, the symbols n and m are the number of alleles sequenced from each of the species, and t is the divergence time since the last common ancestor of the species, scaled as a multiple of the haploid effective population number, N_e. Hence the actual number of generations since the last common ancestor is given by $N_e t$. The parameters θ_s and θ_a are, respectively, the synonymous and nonsynonymous mutation rates, also scaled according to the haploid effective population number as well as scaled according to the number of nucleotides in the sequence. In addition, corrections were made to accommodate the differences in effective population size among different chromosomes. Deleterious mutations are ignored unless they have a nonnegligible chance of becoming polymorphic or fixed. Here, $\theta_s/2$ is the expected number of synonymous mutations that are eligible to become polymorphic or fixed and which occur in the sequence in the entire population in any generation; $\theta_a/2$ is the corresponding quantity for nonsynonymous mutations. The model assumes that all new amino acid mutations have selection coefficients that are scalable by the effective population size. Mutations that are more deleterious disappear immediately in the time scale of the model and are ignored. This results in parameter values θ_a that are smaller than the corresponding scaled site mutation rates and are not easily predictable from θ_s. The divisor of 2 is present because the haploid effective population size equals two times the diploid effective size.

The effects of selection are incorporated in the parameter γ, which is the selection intensity in favor of (if $\gamma > 0$) or against (if $\gamma < 0$) amino acid replacements, again scaled according to the haploid effective population number. The population model has continuous generations, so γ is the Malthusian fitness; it is equivalent to a Darwinian fitness of e^{γ} in a discrete population model (Hartl and Clark 1997). The fitness, γ, maybe thought of as the magnitude of selection affecting amino acid replacements relative to that affecting synonymous substitutions since in this formulation the intensity of selection for optimal codon usage, while known to occur (Hartl et al. 1994; Akashi 1995), is assumed to be small.

A Bayesian Fixed-Effects Model

A Bayesian approach is appropriate for analyzing DPRS tables for a number of coding sequences from the same pair of species because this approach allows a sort of data sharing in which information from all of the coding sequences is used to make inferences about any one of them. This approach also makes use of the theoretical expectations in Eqs. (1)–(4). The theoretical expectations for any single DPRS table include four parameters (θ_s, θ_a, γ, and t) and four observations (K_s, S_s, K_a, and S_a), hence there is no meaningful opportunity for model fitting. However, the divergence time t is a shared parameter among all the sequences and, hence, reduces the number of sequence-specific parameters to fewer than the number of observations. In particular, the data-sharing attribute of the Bayesian approach allows the divergence time (conveniently called a "global parameter") and each of the sequence-specific values of θ_s, θ_a, and γ (conveniently called "local parameters") to be estimated.

The basic idea of Bayesian analysis is to treat the parameters in a model as random variables with some underlying distributions (known as prior distributions) and to treat the data as known constants. The objective is to identify the conditional distribution of

the parameters given the data (the posterior distribution) by means of combining the prior distributions with the observed data using a likelihood function. In hierarchical Bayesian analysis, some parameters (called hyperparameters) are used to define probability distributions for other parameters. Ideally, the greater the number of layers in a Bayesian hierarchy, the less the posterior distribution is dependent on the assumed prior distributions, but in practice hierarchical models rarely go beyond two or three levels (Carlin and Louis 2000).

In the first application of this approach to DPRS tables (Bustamante et al. 2002b), we assumed that, for each coding sequence, the value of γ was a fixed constant but that across loci the distribution of γ was given by a normal distribution with mean μ and standard deviation σ. The hierarchical feature of the model is incorporated by assuming that μ and σ are themselves random variables. Symbolically we can write the posterior distribution of the parameters given the data as $\pi(\gamma, t, \theta, \mu, \sigma|D)$, where now γ and θ are vectors of selection and mutation parameters, respectively. To specify the posterior distribution more explicitly, we can write

$$\pi(\gamma, t, \theta, \mu, \sigma|D) = \frac{P(D|\gamma, t, \theta)f(\gamma|\mu, \sigma)g(\mu|\sigma)h(\sigma)p(\theta)q(t)}{\int\int\int\int\int[P(D|\gamma, t, \theta)f(\gamma|\mu, \sigma)g(\mu|\sigma)h(\sigma)p(\theta)q(t)]d\gamma dt d\theta d\mu d\sigma}$$

where $P(D|\gamma, t, \theta)$ is the posterior probability of the data given and γ, t, θ, and $f(\gamma|\mu, \sigma)$ is the assumed normal distribution of selection coefficients. The other prior distributions are $g(\mu|\sigma)$ assumed to be normal, $h(\sigma)$ assumed to be such that $1/\sigma^2$ is gamma, $p(\theta)$ assumed to be gamma, and $q(t)$ assumed to be uniform. These assumptions are made largely for convenience in computation.

As is typical of Bayesian models with this level of complexity, the expression for $\pi(\gamma,t,\theta,\mu,\sigma|D)$ is analytically intractable. The posterior distribution is nevertheless accessible by means of Markov chain Monte Carlo (MCMC): computer simulations of a Markov chain defined in such a way that the stationary distribution is precisely $\pi(\gamma,t,\theta,\mu,\sigma|D)$ (Gilks et al. 1996; Liu 2001). Essentially, each MCMC trajectory iteratively updates the parameters in either of two ways. One method follows a likelihood criterion (Metropolis et al. 1953) in which a trial value for the new parameter is used to replace the current value if the ratio of posterior probabilities for the trial and present values of the parameter is greater than a uniform random number in [0,1]. Alternatively, parameter values are sampled directly from the conditional distribution. The posterior probabilities of the data are calculated from the relevant parameters, the Poisson distributions for K_s, S_s, K_a, and S_a given in the previous section, and the prior distributions of the parameters. In performing MCMC, it is customary to disregard the first few 10,000 or so iterations (the "burn-in") to minimize possible bias caused by the initial conditions. After the burn-in, periodic sampling from the posterior distribution yields estimates of the parameters and their 95% credible intervals (the Bayesian analog of the 95% confidence interval).

A Bayesian Random-Effects Model

It is of course unrealistic to suppose that all new mutations that are eligible to become either polymorphic or fixed will have the same selection coefficient. To make this aspect of the Bayesian model more realistic, we suppose that the selective effects of each new mutation in a gene conform to some statistical distribution. Since γ measures the selection intensity in terms of the scaled Malthusian parameter, it is reasonable to suppose that the distribution of γ among the new mutations at a locus is a normal distribution. The corresponding distribution of Darwinian fitnesses is then log-normal.

Accordingly, we have implemented a Bayesian model in which, for the ith coding sequence, the distribution of selection intensities among newly arising mutations that are eligible to become either polymorphic or fixed is a normal distribution with mean γ_i and standard deviation σ_w. Technically, σ_w is included in the model with a uniform prior distribution. Here, γ_i is a local parameter that differs for each gene in the data set but σ_w is a global parameter whose value depends on the totality of the data. In effect this means that the within-locus variation in selection intensity is the same for each locus. Similarly as in the fixed-effects model described earlier, we assume that the local γ_i values are drawn from a normal distribution with mean μ and between-locus standard deviation σ_b.

Application to Data From *Drosophila*

The random-effects model was originally implemented using a set of 72 coding sequences obtained from GenBank that included data on polymorphisms in natural populations of *D. simulans*. These data are summarized in the Supplementary Information. Among the coding sequences, the

number of sequenced alleles from *D. simulans* ranged from 4 to 70, with an average of 10.5. Nucleotide divergence between *D. simulans* and *D. melanogaster* was inferred from the reference sequence of *D. melanogaster* (Adams et al. 2000), hence the sample size for *D. melanogaster* is one. The use of a single reference sequence does not compromise the analysis because the theoretical expectations in Eqs. (1)–(4) are unbiased estimators that take the differing sample sizes into account (Sawyer and Hartl 1992). In fact, the use of a single reference has the advantage that it avoids minor complications that otherwise arise because of possible different effective population sizes between *D. simulans* and *D. melanogaster* (Akashi 1995).

The counts (K_s, S_s, K_a, and S_a) at each locus are used in the posterior density as observations of independent Poisson variates whose parameters are given by Eqs. (1)–(4). Each of these is a count of different types of mutations corresponding to the different terms in these equations. For example, the counts for polymorphisms are sums of two terms, one for each species, representing mutations that have caused a population-wide polymorphism in that species that is also polymorphic in the sample. This implies that a site that is polymorphic in both species should be counted as two polymorphic sites, not one, although this does not apply in our case because only one *D. melanogaster* sequence is used. However, sites that have more than two nucleotides segregating in the same species should be counted as more than one polymorphic site. For fixed differences, the counts are sums of four terms that include, for each species, one term representing population-wide monomorphisms and one term for sample monomorphisms that occur by chance even though the site is polymorphic in the population as a whole.

Second, the counts should be codon based rather than nucleotide based, even for silent sites. This is necessary to tabulate properly polymorphic codon positions that have two or more polymorphic sites but only two segregating codons. This provision is necessary because the theory implicitly stipulates that no more than one mutational event can occur per codon. In real data, a significant number of codon positions can have more than two segregating codons, and then one needs to make inferences based on parsimony about how many events and of what type have occurred. The following accounting rules seem to capture most situations. For any codon, (1) if the set of segregating codons in species 1 is nonoverlapping with the set of segregrating codons in species 2, add 1 to the count of apparent population (and therefore sample) fixed differences; (2) if species 1 has n_1 distinct codons segregating and species 2 has n_2 distinct codons segregating, then add n_1+n_2-2 to the count of sample polymorphisms.

In applying the random-effects model to the DPRS data, we initially found that the Markov chain did not converge, or did so excessively slowly. This may have been caused by the addition of the extra global parameter σ_w with no reduction in the number of local parameters. The output for various runs suggested that the main reason for poor convergence was that values of θ_r (the scaled frequency of replacement mutations eligible for polymorphism or fixation) could apparently undergo trajectories that balanced off γ (the scaled selection intensity). This behavior can be rationalized by the argument that an excess of replacement mutations can be caused either by a stronger intensity of positive selection or by a higher frequency of positive mutations.

From runs of the fixed-effects model (Bustamante et al. 2002b), we noticed that about 80% of the coding sequences in the DPRS tables had values of $\theta_r/2\theta_s$ that were reasonably similar and smaller than approximately $\frac{1}{4}$ (actually, 0.28). This suggested a strategy of regarding $\theta_r/2\theta_s$ as a fixed constant and including it as another global parameter, which reduces the number of local mutation parameters by half. In effect, fixing $\theta_r/2\theta_s$ requires each coding sequence to have approximately the same fraction of replacement mutations, relative to synonymous mutations, that are eligible for polymorphism or fixation. Technically, a global parameter, $Q = \theta_r/2\theta_s$, is included in the model with a gamma prior distribution and $\theta_r = 2Q\theta_s$. Among the 72 genes, 14 were excluded because $\theta_r/2\theta_s > 0.28$ and two additional loci were excluded because they appeared to be suspicious for other reasons. These genes, most of which have exceptionally high rates of amino acid replacement, were *Acp32CD*, *Acp33A*, *Acp36DE*, *Acp53Ea*, *Acp62F*, *Acp63F*, *Acp76A*, *Acp98A*, *anon1A3*, *anon1E9*, *anon1G5* (*cav*), *AP-50*, *bnb*, *ct*, *Hsp70Aa*, and *Osbp*. This list includes eight genes for male accessory gland proteins (*Acp**) and three genes encoding proteins of unknown function (*anon**). At least some of these genes were included in population studies precisely because they were known or suspected to be undergoing rapid amino acid replacement and therefore were strong candidates for positive selection (Schmid and Tautz 1997; Swanson et al. 2001).

Excluding the outliers with $\theta_r/2\theta_s > 0.28$ reduced the number of individual DPRS tables to 56. When $\theta_r/2\theta_s$ was treated as a constant and fitted as a global parameter, the resulting Markov chain for the random-effects model converged and mixed very well. The run had 1,000,000 iterations after an initial burn-in of 10,000 iterations. After the burn-in, every 10th iteration was used as a sample. The 100,000 samples were split into ten consecutive subchains of 10,000

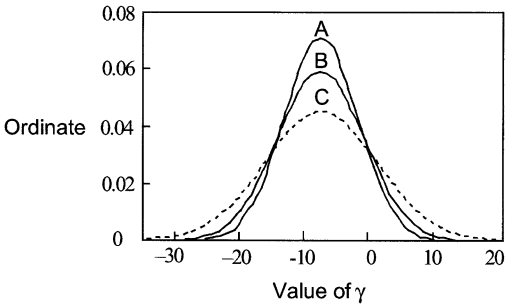

Fig. 1. Posterior distributions of the scaled selection intensities ($\gamma = N_e s$). Curve A is the distribution of mean selection intensities across loci. Curve B is the distribution of effects within coding sequences; the mean differs from one gene to the next, but for purposes of illustration the grand mean across loci ($\gamma = -7.31$) is used. Curve C is the combined distribution taking into account the variance within and between loci. The standard deviations are 5.69 (A), 6.79 (B), and 8.86 (C).

samples each. A standard criterion for convergence based on the subchains is the Gelman–Rubin coefficient (Gelman et al. 1997), which varied in the range 0.99998–1.00272 for the global parameters and the local selection parameters, γ_i. Another criterion is R^2, which is the proportion of the variability of the sample values that is attributable to subchain means. This varied in the range 0.00068–0.025.

Distribution of Selection Coefficients

For the random-effects model, Fig. 1 shows normal distributions of the selection coefficients for amino acid replacements based on polymorphisms in 56 genes from *D. simulans* and divergence from a reference sequence of *D. melanogaster*. In all cases the distribution is centered on a mean selection intensity of $\gamma = -7.31$ estimated from the MCMC, with 95% credible interval (−20.67, −0.34). We emphasize that γ is the selection intensity of all newly arising amino acid replacement mutations that are eligible to become polymorphic or fixed, not that of all newly arising mutations. Nevertheless the negative value of γ supports the traditional intuition that most amino acid replacements are deleterious.

In Fig. 1, distribution A is the narrowest. It is the distribution of mean gamma values, γ_i, among loci. Its mean is $\gamma = -7.3$; its standard deviation, σ_b, is 5.69. In contrast, distribution B is the distribution of selection intensities at a representative locus whose mean is $\gamma = -7.3$. The standard deviation of this distribution is $\sigma_w = 6.79$. Finally, distribution *C* is the estimated overall distribution of selection intensities within and between loci and its standard deviation, which equals $\sqrt{(\sigma_w^2 + \sigma_b^2)}$, is 8.86.

Figure 2 depicts the mean selection intensity for each of the 56 genes, ranked by magnitude from smallest to largest, and their 95% confidence intervals. The sampling distribution of the selection intensities is sufficiently symmetrical that the values for the medians are close to those of the means and the 95% confidence and 95% credible intervals were nearly identical, except that the latter were more negative for the first few values of γ (data not shown). Although most of the confidence intervals in Fig. 2 overlap 0, the means for 51 of the genes are negative. (The five genes with positive mean γ are *otu*, *ase*, *Acp29AB*, *Rel*, and *mei-218*.) The most negative mean γ is that of the gene *Pgm* with $\gamma = -16.0$. Nevertheless, the variance of distribution A in Fig. 1 is sufficiently large that, even for *Pgm*, about 1% of new amino acid replacements have selection coefficients that are positive. (The average for this quantity across all 56 genes is 19.4%.)

Evidence for Positive Selection of Amino Acid Replacements

In the MCMC runs, we also monitored what fraction of the newly arising amino acid replacements eligible to become polymorphic or fixed was beneficial, what fraction of sample polymorphic amino acid replacements is estimated to be beneficial, and, finally, what fraction of fixed amino acid replacements in the population was beneficial. These values are implicit in the normal distribution of selection intensities stipulated in the random-effects model with standard deviation $\sigma_w = 6.79$ (curve B in Fig. 1), where the mean for each gene is given by the center value in Fig. 2.

The results are summarized in Fig. 3, in which the fraction of positively selected mutations in each class is denoted P(+). For most genes, the majority of new mutations among those eligible to become polymorphic or fixed are deleterious. Averaged across MCMC runs, the fraction of beneficial new mutations (open circles) ranges from 1% (for *Pgm*) to 62% (*Rel*), with an outlier at 90% (*mei-218*). The average for all loci is 19.4%. Among the replacement polymorphisms in the data, the estimated average proportion that is beneficial is almost uniformly distributed over the loci, ranging from 2.9% (for *vermilion*) to 98% (for *mei-218*) and averaging 46.9%.

One of the principal results of this study is that, judging from the results of the Bayesian random-effects model, the majority of fixed amino acid replacements between *D. simulans* and *D. melanogaster* are positively selected. This feature of the analysis is shown by the filled circles in Fig. 3. With the exception of the outlier at 34% positively selected (*kl-5*), the fraction of fixed amino acid replacements that are beneficial ranges from 72% (*sog*) to 99.92% (*mei-218*) and averages 93.2%.

Among amino acid replacements that are fixed, the average selection intensities are shown by rank in Fig.

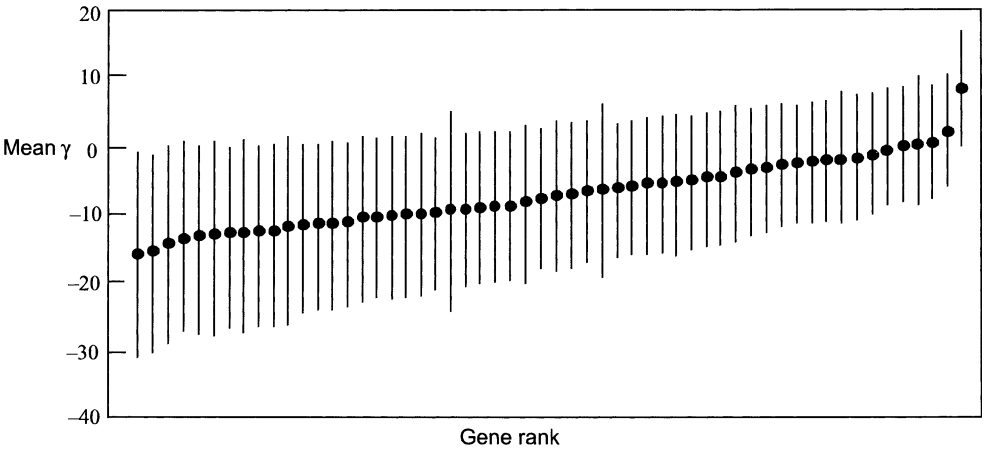

Fig. 2. Mean and 95% credible interval of the selection intensity among newly arising amino acid replacements, ranked in increasing order of the mean. In the model, the only mutations that are relevant are ones that have a chance to become polymorphic or fixed, hence very severely deleterious mutations are ignored.

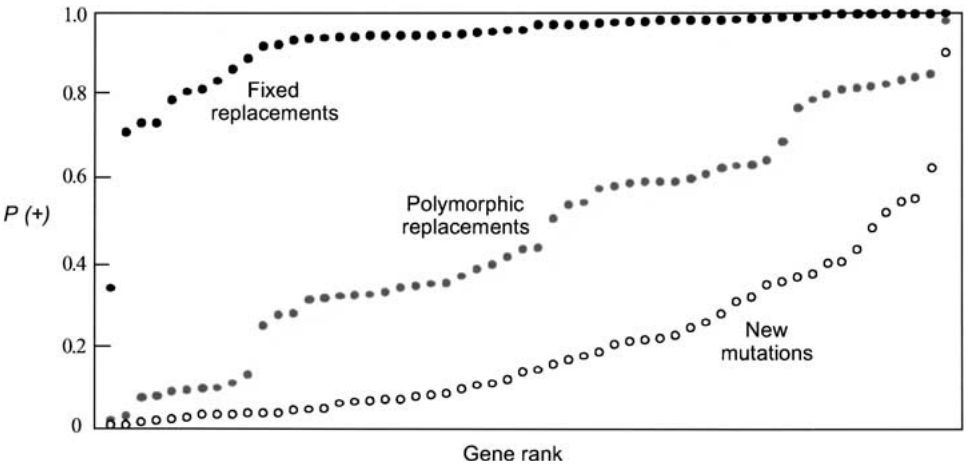

Fig. 3. Proportion of amino acid replacement mutations that are positively selected, denoted P(+), among new mutations (open circles), sample polymorphisms (shaded circles), and fixed differences (filled circles).

4. There is one outlier, the Y-linked gene for dynein ATPase corresponding to fertility factor, *kl-5*, for which the estimated average selection intensity for fixed mutations is −0.38. All of the others are positive, and perhaps unexpectedly, none are smaller than +2.0. The range is from +2.1 (for the X-linked gene *vermilion*) to +9.4 (for *Rel*), with an overall mean (excluding *kl-5*) of +5.1. For *Pgm*, which has the smallest fraction of newly arising amino acid replacements that are beneficial, the proportion of fixed mutations that are positively selected is 94%, and among these the average selection intensity is +3.6.

While large relative to the effective population size, the level of positive selection is small in absolute terms. As an order-of-magnitude approximation, we may take N_e for *D. simulans* as approximately 10^6 (Akashi 1995), in which case a value of $\gamma = +5.1$ implies a conventional overall average selection coefficient of 5.1×10^{-6}.

Polymorphism and Fixation of Slightly Deleterious Mutations

The high frequency of fixed replacements driven by positive selection shown in Fig. 3 has the counterpart that deleterious mutations are usually not fixed, even for genes in which a substantial fraction of segregating replacement polymorphisms are deleterious. These results are shown in Fig. 5, where P(−) denotes the proportion of mutations in each class that are deleterious. The genes are ranked according to the proportion of fixed replacements that are deleterious (open circles). The outlier with deleterious fixations at 66% is again the Y-linked fertility factor, *kl-5*. Without this exception, the proportion of fixed replacement differences that are deleterious ranges from 0.08% (*mei-218*) to 28.5% (*sog*), with an average of 5.7%. While this is by no means negligible, neither does it suggest that a large fraction of

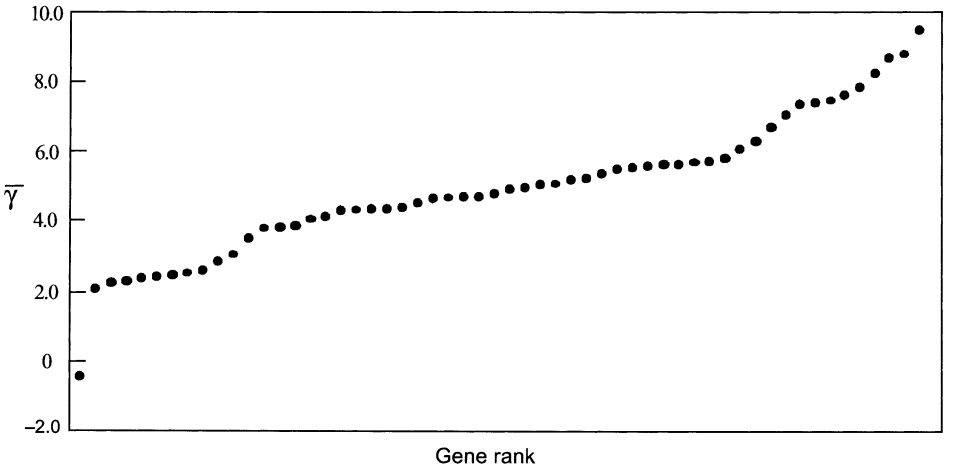

Fig. 4. Average scaled selection intensity ($\gamma = N_e s$) among fixed mutations, ranked in order. The negative outlier is the Y-linked fertility factor, *kl-5*.

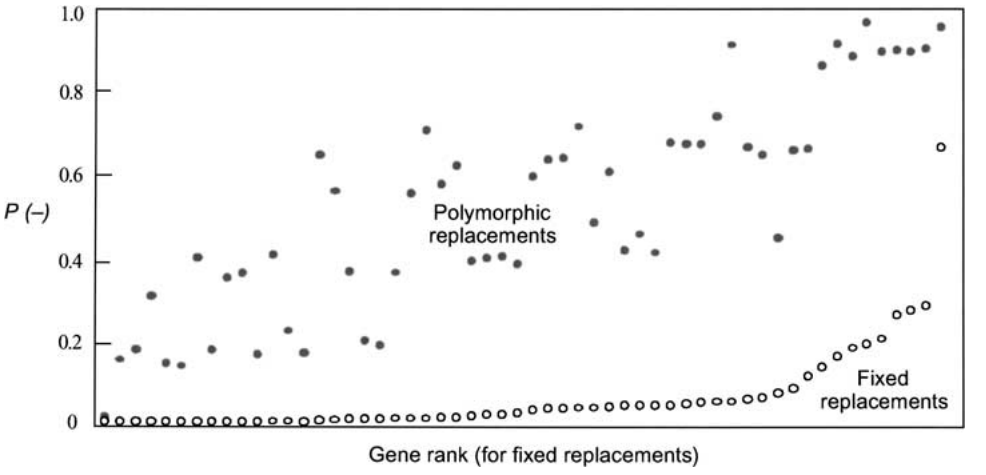

Fig. 5. Proportion of amino acid replacement mutations that are deleterious, denoted P(−), among fixed replacements (open circles) and sample polymorphisms (shaded circles), ranked by the fixed replacements. The outlier with P(−) = 0.66 for fixed replacements is again *kl-5*.

fixed amino acid replacements are slightly deleterious (Ohta 1992).

Although the correlation between the fraction of polymorphic amino acid replacements that are deleterious (filled circles) and the fraction of fixed replacements that are deleterious is not perfect, it is large ($r = 0.75$, excluding *kl-5*) and highly significant. For polymorphisms, the range of P(−) is from 2.4% (*mei-218*) to 97.1% (*vermilion*), with an average of 53.1%. The Bayesian random-effects model therefore implies that about half of all polymorphic amino acid replacements are deleterious (with a very large variance from locus to locus) but that only a relatively small fraction of these polymorphisms is destined to become fixed between species. In contrast, the Bayesian fixed-effect model (Bustamante et al. 2002b) assumes that all new weakly selected mutations have the same selection coefficient at the same locus. In this model, 99.96% of new amino acid mutations in *Drosophila* were positively selected, and hence the same percentage (99.96%) among replacement sample polymorphism and among replacement fixed differences.

The Efficacy of "Wedging"

The random-effects model is a particular model that specifies aspects of the mutational process that are in reality unknown. It assumes that, at a particular locus, the distribution of selective effects of mutations eligible to become polymorphic or fixed is Gaussian, that the standard deviation of this distribution is the same for all loci, and that the distribution of fitness effects of new mutations remains the same through time. Other caveats intended to discourage a too literal interpretation of the model are examined in the next section.

Nevertheless, one consistent finding is pervasive positive selection for amino acid replacements that

become fixed between species (Fig. 3). This is true of all loci with the exception of the Y-linked gene, *kl-5*, which might be expected to be an outlier owing to the heterochromatic nature of the Y chromosome, the complete lack of opportunity for recombination, and the small number of functional genes on this chromosome (Charlesworth and Charlesworth 1997). For the other genes, the average proportion of amino acid replacement mutations that have a chance to become polymorphic or fixed is 19.7%. This fraction is enriched by a factor of 2.4 among polymorphic replacements to become 47.7%, which is enriched further by a factor of 2.0 among fixed replacements to become 94.3%. Relative to the frequency of new mutations that are beneficial, the enrichment among fixed replacements is by a factor of 4.8.

The effectiveness of selection in seizing upon a minority of favorable mutations is seen nearly across the board for each gene individually. The factor of enrichment for beneficial polymorphisms ranges from lows of 1.1 (*mei-218*) and 1.3 (*Rel*) to highs of 9.0 (*Pgm*) and 9.6 (*Hsc70-4*). The enrichment of beneficial fixations relative to beneficial polymorphisms ranges from lows of 1.0 (*mei-218*) and 1.2 (tied across *Acp29AB*, *tra2*, *Cen190*, *otu*, *ase*, *Gel Rel*, *mth*, and *nos*) to highs of 11.2 (*Pgm*) and 27.9 (*vermilion*). Relative to new mutations, the increase in P(+) among fixed replacement differences ranges from 1.1 (*mei-218*) and 1.6 (*Rel*) to 70.4 (*vermilion*) and 101 (*Pgm*). As might be expected, there are significant correlations between these factors of enrichment across genes—0.87 for P(+) between polymorphisms and P(+) for new mutations, 0.75 for P(+) between fixations and P(+) for polymorphisms, and 0.48 for P(+) between fixations and P(+) for new mutations. While highly significant ($P < 0.01$ in all cases), these correlations are by no means perfect, which reflects the fact that the inferred underlying distribution of the selective effects of new mutations differs from one gene to the next.

The random-effects model lends emphasis to the strong tendency for natural selection to seize favorable mutations to become polymorphic or fixed. In his notebooks, Darwin frequently used the term "wedging" to describe the process of competition by which one species displaced another (Gould 1989). Figure 3 shows a sort of wedging at the molecular level, in which natural selection tends to displace deleterious new mutations from the set of mutations that become polymorphic and then to displace deleterious polymorphisms from the set of mutations that become fixed.

Caveats

While the Bayesian random-effects model has the virtues of specificity, concreteness, and explicit assumptions, like other theoretical models it ignores many potential complications. The selection model is essentially haploid selection, which for a diploid means additive effects of alleles, with the Malthusian fitness of heterozygous genotypes equal to the mean of the fitnesses of the corresponding homozygous genotypes. Since the inferred selection intensities are relatively small, this assumption seems justified. However, the model in its present form allows no scope for heterozygote superiority or such departures from constant fitness as frequency-dependent selection or fluctuating selection intensities (Gillespie 2000).

The model also ignores demographic factors that may affect the fate of mutant alleles. It assumes a constant effective population number through time, and adjustments would have to be made for populations undergoing growth or shrinkage. There is no geographical population structure in the model and no migration or extinction and recolonization of local populations. As a practical matter this means that the sampling strategy should be given careful consideration in generating polymorphism–divergence data. Since almost all real populations have some sort of geographical subdivision, it is important to sample widely in order to include diversity across the population as a whole. Otherwise, the estimates of polymorphism and divergence will be biased. This is perhaps especially important in inbred populations, where the local heterozygosity may be much reduced (Wakeley 2000).

The random-effects model also invokes a simple model of mutation in which the distribution of selection intensities among amino acid replacement mutations that may become polymorphic or fixed is assumed to be homogeneous in time. This assumption would be invalidated if each successive polymorphism or fixation changes the distribution of selection intensities among subsequent new mutations. This model also assumes that the fitness effects of mutations are additive across sites in a gene, hence there is no possibility of epistatic interactions among mutations. In particular, the model in its present form cannot handle compensatory mutations (Hartl and Taubes 1996; Stephan 1996), in which the fixation of one or more neutral or slightly deleterious mutations changes the mutational spectrum in such a way that mutations with beneficial epistatic interactions become possible.

The model also assumes independence between nucleotide sites at each locus when in fact the nucleotides within a gene are more or less tightly linked according to the local rate of recombination. Preliminary data from simulations suggest that the assumption of independence results in a slight negative bias in the estimate of the selection intensities, hence the inference of pervasive positive selection is strengthened. The effects of tight linkage nevertheless warrant further investigation. For example, with

frequent positive selection, tight linkage results in a phenomenon of interference selection, which, compared with neutrality, decreases the level of polymorphism, increases the proportion of rare variants, and promotes linkage disequilibrium (Comeron and Kreitman 2002). It is not clear what effects interference selection would have on the inference of pervasive positive selection that emerges from the random-effects analysis.

Independence between loci is also assumed in our analysis. This assumption would be invalidated in regions of reduced recombination in which there are selective sweeps of strongly favorable mutations (Maynard Smith and Haigh 1974; Galtier et al. 2000; Nurminsky 2001). These reduce the effective population size in the region so that the species no longer has a unique effective population size overall, but, rather, a different effective population size for each gene or genomic region. Also, there may be selective sweeps which, depending on the tightness of linkage across the region, cause interference-like effects called trafficking in which fixation of a single haplotype is delayed until a recombination event brings the two variants together on the same chromosome (Kirby and Stephan 1996; Kim and Stephan 2000).

One of the unexpected and surprising features of the random-effects model is that it apparently cannot easily accommodate the fixation of a large number of deleterious mutations. As shown in Fig. 5, except for the outlier *kl-5*, the proportion of fixed mutations with $\gamma < 0$ is not larger than 28%, and across all genes it averages just 5.7%. Even if the roles of the synonymous and replacement sites are interchanged (that is, replacement mutations are treated as neutral and the synonymous sites as under selection), the model implies that about 50% of the fixed differences (in this case, synonymous differences) are positively selected (data not shown). The basis of this behavior is unclear, but it may relate to the assumption that the prior distribution of fitness effects is Gaussian. In this case the right-hand tail of effects decreases as e^{-x^2}. It is possible that a distribution with a more rapidly decreasing right-hand tail, such as e^{-x^3}, would increase the relative proportion of fixations that are mildly deleterious. It would be of great interest to explore alternative models that would allow more mildly deleterious mutations to be fixed. On the other hand, there are no generally accepted criteria for judging whether, as regards to their fit to the data, the output of one Bayesian model is significantly better than that of another model.

Comparison with Heuristic Analyses

The results of the random-effects model are generally consistent with two recently published heuristic analyses of DPRS tables from species of *Drosophila* closely related to *D. melanogaster*, but they differ in some of the details. For example, from their analysis of polymorphism and divergence in *D. simulans* and *D. yakuba*, Smith and Eyre-Walker (2002) deduce that about 45% of the amino acid replacements between these species have been driven by positive selection. Their data suggest that these species have undergone one amino acid replacement every 20 years (~200 generations), or about 600,000 substitutions altogether, of which 270,000 were driven by selection. As noted, the random-effects model does not imply that about 55% of amino acid replacements are neutral or nearly neutral. In this model, at least 70% of the amino acid fixations of an autosomal or X-linked origin are driven by positive selection, and the average is 94.3%.

Fay et al. (2002) have analyzed data from 45 genes in *D. melanogaster* and *D. simulans* from a somewhat different perspective and have come to a somewhat different conclusion. While they note evidence for positive selection in the data as a whole, they attribute most of the positive selection to 11 genes (*Acp26Aa*, *Acp29AB*, *anon1A3*, *anon1E9*, *anon1G5*, *ci*, *est-6*, *Ref2P*, *Rel, tra*, and *Zw*) and regarded the remaining 34 genes as evolving essentially neutrally with respect to amino acid replacements. Certain genes were excluded from our analysis because of their large values of $\theta_r/2\theta_s$, including three in the above list (*anon1A3*, *anon1E9*, *anon1G5* [*cav*]). In the analysis of the other genes, we found no indication that they could be split into two groups, one accounting for most of the positive selection and the other in which fixations are largely neutral. Figures 3 and 4 show that the random-effects model ascribes most fixations of most genes to positive selection and that the average selection intensity of most fixations is greater than 2.

Finally, we note that the heuristic analyses of polymorphism and divergence as well as our Bayesian approach all assume that synonymous nucleotide substitutions are nearly neutral. The amino acid polymorphisms and replacements are interpreted on that basis. If it turns out that this assumption is not justified, then the apparent evidence for gene-specific (Fay et al. 2002), widespread (Smith and Eyre-Walker 2002), or nearly pervasive (our results) positive selection of amino acid replacements will have to be reevaluated.

Acknowledgments. This work was financially supported by National Institutes of Health Grants GM60035 and GM65169 (D.L.H.), National Science Foundation Grant DMS-0107420 (SAS), and fellowships from the Natural Sciences and Engineering Council of Canada (R.J.K.) and the Marshall-Sherfield fund (C.D.B.)

References

Adams MD, Celniker SE, Holt RA, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139:1067–1076

Bustamante CD, Nielsen R, Hartl DL (2002a) A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. Mol Biol Evol 19:110–117

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002b) The cost of inbreeding in *Arabidopsis*. Nature 416:531–534

Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis. Chapman & Hall, London

Charlesworth B, Charlesworth D (1997) Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res 70:63–73

Comeron JM, Kreitman M (2002) Population, evolutionary and genomic consequences of interference selection. Genetics 161:389–410

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fay JC, Wyckoff GJ, Wu C-I (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415:1024–1026

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

Galtier N, Depaulis F, Barton NH (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–987

Gelman A, Carlin JS, Stern HS, Rubin DB (1997) Bayesian data analysis. Chapman & Hall, London

Gilks R, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London

Gillespie JH (2000) Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155:909–919

Gould SJ (1989) Tires to sandals: As we strive to understand nature, do we seek truth or solace? Nat Hist 98:8–15

Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland, MA

Hartl DL, Taubes CH (1996) Compensatory nearly neutral mutations: Selection without adaptation. J Theor Biol 182:303–309

Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138:227–234

Hudson RR (1990) Gene geneologies and the coalescent process. In: Futuyma D, Antonovics J (eds) Oxford surveys in evolutionary biology,. Oxford University Press, Oxford, pp 1–44

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics 136:1329–1340

Kim Y, Stephan W (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 155:1415–1427

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kirby DA, Stephan W (1996) Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics 144:635–645

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York

Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genet Res 23:23–35

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1091

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Nurminsky DI (2001) Genes in sweeping competition. Cell Mol Life Sci 58:125–134

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176

Schmid KJ, Tautz D (1997) A screen for fast evolving genes from *Drosophila*. Proc Natl Acad Sci USA 94:9746–9750

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024

Stephan W (1996) The rate of compensatory evolution. Genetics 144:419–426

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci USA 98:7375–7379

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Templeton AR (1998) Nested clade analysis of phylogeographical data: Testing hypotheses about gene flow and population history. Mol Ecol 7:381–397

Templeton AR (2002) Out of Africa again and again. Nature 416:45–51

Wakeley J (2000) The effects of subdivision on the genetic divergence of populations and species. Evol Int J Org Evol 54:1092–1101

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573