The Genomic Rate of Adaptive Amino Acid Substitution in *Drosophila*

Nicolas Bierne[1] *and Adam Eyre-Walker*

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton, UK

The proportion of amino acid substitutions driven by adaptive evolution can potentially be estimated from polymorphism and divergence data by an extension of the McDonald-Kreitman test. We have developed a maximum-likelihood method to do this and have applied our method to several data sets from three *Drosophila* species: *D. melanogaster*, *D. simulans*, and *D. yakuba*. The estimated number of adaptive substitutions per codon is not uniformly distributed among genes, but follows a leptokurtic distribution. However, the proportion of amino acid substitutions fixed by adaptive evolution seems to be remarkably constant across the genome (i.e., the proportion of amino acid substitutions that are adaptive appears to be the same in fast-evolving and slow-evolving genes; fast-evolving genes have higher numbers of both adaptive and neutral substitutions). Our estimates do not seem to be significantly biased by selection on synonymous codon use or by the assumption of independence among sites. Nevertheless, an accurate estimate is hampered by the existence of slightly deleterious mutations and variations in effective population size. The analysis of several *Drosophila* data sets suggests that approximately 25% ± 20% of amino acid substitutions were driven by positive selection in the divergence between *D. simulans* and *D. yakuba*.

Introduction

It has long been appreciated that a way to test the neutral theory of molecular evolution is to compare polymorphism and substitution data (Kimura and Ohta 1971; Chakraborty, Fuerst, and Nei 1978; Skibinski and Ward 1982; Kimura 1983). To completely reconcile the evolutionary parameters shared in common by the two types of data, one needs to refer to a neutral control, on which selection is very unlikely (Kimura 1983). McDonald and Kreitman (1991) proposed the use of synonymous (silent) mutations that do not result in a change in the protein as a neutral reference to infer the direction of the selection acting on nonsynonymous (replacement) mutations. In their method, numbers of synonymous polymorphisms (P_s) and substitutions (D_s) and numbers of nonsynonymous polymorphisms (P_n) and substitutions (D_n) are compared in a contingency table. If nonsynonymous mutations are either neutral or strongly deleterious, the ratio D_n/D_s should be equal to the ratio P_n/P_s (Kimura 1983; McDonald and Kreitman 1991; Sawyer and Hartl 1992). Under the assumption that the selective constraint has not increased owing to an increase in effective population size, an excess of nonsynonymous substitutions relative to polymorphisms ($D_n/D_s > P_n/P_s$) implies the fixation of advantageous mutations. The beauty of the McDonald-Kreitman (MK) approach is that neutral and selected sites are interspersed with each other throughout the gene. They, therefore, have the same evolutionary history and sampling (i.e., a shared phylogeny and the same effective population size.) The test is, therefore, free from specific population genetics and can be performed using a simple chi-square or *G*-test of independence. It was anticipated as soon as the MK test was developed that its application to many genes would help estimate the importance of positive Darwinian selection in evolution (review in Brookfield and Sharp [1994], Kreitman and Akashi [1995], and Moriyama and Powell [1996]). However, only recently has sufficient data become available to make this possible.

The rate of amino acid substitution can be estimated from the MK test (Charlesworth 1994; Akashi 1999; Fay, Wycoff, and Wu 2001; Smith and Eyre-Walker 2002). The number, *a*, and the proportion, α (= a/D_n), of amino acid substitutions driven by positive selection in a gene can be estimated by the following equations:

$$a = D_n - D_s \frac{P_n}{P_s}, \qquad \alpha = 1 - \frac{D_s P_n}{D_n P_s} \tag{1}$$

Levels of nucleotide diversity and amino acid divergence are often low such that most genes have only a few polymorphic sites and a few nonsynonymous substitutions. Single gene estimates, therefore, have large variances. For this reason, data needs to be combined across genes. Three methods have been used to do this, but each has its potential problems. First, data have simply been summed across genes. Using this method Fay, Wycoff, and Wu (2001) estimate that approximately 35% of the amino acid substitutions between humans and Old World monkeys were a consequence of positive selection. However, this method is likely to be biased if the selective constraint, as measured by the ratio P_n/P_s, is correlated with gene diversity; that is, there are slightly deleterious mutations segregating, and the effective population size varies across the genome. There is evidence that some nonsynonymous mutations are slightly deleterious in some species because nonsynonymous polymorphisms tend to segregate at lower frequencies than synonymous mutations in *Drosophila melanogaster* (Akashi 1996; Fay, Wycoff, and Wu 2002) and humans (Cargill et al. 1999; Fay, Wycoff, and Wu 2001). In addition, the correlation between local recombination rates and gene diversity observed in both of these species (*D. melanogaster* [Begun and Aquadro 1992] and humans [Nachman et al. 1998]) suggests that the effective population size varies across the genome. To overcome the potential bias, Smith and Eyre-Walker (2002) suggested a second method of combining

[1] Present address: Laboratoire Génome, Populations, Interactions, Adaptation, UMR 5171 CNRS-UMII-IFREMER, Station Méditerranéenne de l'Environnement Littoral, Sète, France.

Key words: Adaptive evolution, McDonald-Kreitman test, amino acid substitutions, *Drosophila*.

E-mail: a.c.eyre-walker@sussex.ac.uk.

Mol. Biol. Evol. 21(7):1350–1360. 2004
doi:10.1093/molbev/msh134
Advance Access publication March 24, 2004

data across genes. Using their method, they estimated that approximately 45% of amino acid substitutions were driven by positive selection in the divergence between *D. simulans* and *D. yakuba*. However, they had to exclude genes that had little or no polymorphism. Furthermore, there was the possibility that the evidence of adaptive evolution was an artifact, produced by the fixation of slightly deleterious mutations in a smaller ancestral population (McDonald and Kreitman 1991; Fay, Wycoff, and Wu 2002; Eyre-Walker 2002). Finally, Bustamante et al. (2002) have recently developed a method to combine data across genes by a hierarchical Bayesian analysis. Following the theoretical work of Sawyer and Hartl (1992), Bustamante et al. (2002) looked at the problem slightly differently: in their method, the parameter of interest is the "average" selection intensity (i.e., selection scaled by the population size, N_es) acting on nonsynonymous mutations that contribute to polymorphism and substitution. This model implicitly assumes that all the nonsynonymous polymorphisms and substitutions share the same weak strength of selection. Unfortunately, the number of parameters that can be estimated from an MK table is restricted, and, unless further information can be incorporated in the model, it seems difficult to estimate jointly the rate of adaptive substitutions and the fitness effect of adaptive mutations.

Here, we present a simple method to estimate the rate of adaptive amino acid substitution. The method retains the benefits of the MK test because it compares synonymous and nonsynonymous sites that have the same evolutionary history and sampling scheme, but it combines data from all genes, including those that have little or no polymorphism. The conditions under which the inference is valid are investigated by applying the method to several data sets in *Drosophila*, with the aim of disentangling the effect of demography and adaptive evolution.

Materials and Methods

Data

As a basis for comparison, we first reanalyzed the same data set as Smith and Eyre-Walker (2002). It is composed of 35 genes with polymorphism data in *Drosophila simulans* and divergence data between *D. simulans* and *D. yakuba*. One aim of the present work was to perform a comparative analysis among different species that may have different demography. To do this, we compiled polymorphism data on coding sequences from 88 genes: 75 genes in *D. simulans* and 57 genes in *D. melanogaster*. Polymorphism data were available in both species for 44 genes. A *D. yakuba* sequence allowed us to separate substitutions within each lineage for 22 out of these 44 genes. In addition to coding sequences, intron polymorphism data were available for 34 loci in *D. simulans* and 38 loci in *D. melanogaster*. Gene names, polymorphism, and substitution data are provided in Supplementary Material online available on the MBE Web site. The raw data (i.e., alignment files) are available from the authors upon request.

For the closely related species *D. simulans* and *D. melanogaster*, polymorphisms and substitutions (P_n, P_s, D_n, and D_s) were counted following McDonald and Kreitman's (1991) recommendations using DnaSP (Rozas and Rozas 1999), having verified that a correction for multiple hits was unnecessary. When divergence with the more distantly related species *D. yakuba* was used, substitutions were calculated using the method of Goldman and Yang (1994) as implemented in the program codeml of the PAML package (Yang 1999).

To test the effect of some factors on the assumptions of our model, we have used the largest data set (75 genes) and have split it into two halves according to the factor under consideration. Linkage disequilibrium was calculating as the average of r^2 (Hill and Robertson 1968) over pairwise comparisons. Genes with less than six informative pairwise comparisons were removed from the data set before it was split. Codon usage bias was measured by the frequency of optimal codons (*Fop* [Ikemura 1985]). Finally, we investigated the consequences of excluding rare variants by removing singletons from our polymorphism data.

Model

Following the original MK test, we use a simple model that does not rely on a specific population genetic model. The expected numbers of synonymous ($\hat{P}_{si}$) and nonsynonymous ($\hat{P}_{ni}$) polymorphisms segregating in a sample of n_i sequences of a locus i from a population are

$$\hat{P}_{si} = \Theta_i L_i \qquad \hat{P}_{ni} = \omega_i \Theta_i L_i \tag{2}$$

where Θ_i is a measure of the synonymous diversity (i.e. the average number of synonymous polymorphisms per codon), L_i is the length of the sequence in codons, and ω_i is the nonsynonymous to synonymous diversity ratio ($= P_{ni}/P_{si}$).

We are assuming that all synonymous mutations are neutral and that all nonsynonymous mutations are either strongly deleterious, neutral, or strongly advantageous. "Strongly," is assumed to mean that advantageous mutations contribute little to polymorphism, although they may contribute substantially to the divergence between species. This is not an unrealistic assumption; at most, an advantageous mutation will contribute twice as much heterozygosity during its lifetime as a neutral variant (Kimura 1983). As we show in the *Appendix*, ignoring the contribution of advantageous mutations to polymorphism leads to an underestimate of the adaptive substitution rate; however, this bias is negligible when N_es is greater than 20. Under this model, the expected numbers of synonymous ($\hat{D}_{si}$) and nonsynonymous ($\hat{D}_{ni}$) substitutions at locus i are

$$\hat{D}_{si} = \lambda_i L_i \qquad \hat{D}_{ni} = (\omega_i \lambda_i + \eta_i L_i) \tag{3}$$

where λ_i is the synonymous substitution rate per codon and η_i is the number of adaptive substitutions per codon at locus i. Here, the ω ratio (P_n/P_s) gives us an estimate of the proportion of nonsynonymous mutations that are neutral; this estimate, used in combination with λ, allows one to predict the number of neutral nonsynonymous substitutions. The difference between this prediction and the actual number of nonsynonymous substitutions is then interpreted as being caused by adaptive substitutions. We can also write $\hat{D}_{ni}$ as

$$\hat{D}_{ni} = \frac{\omega_i \lambda_i L_i}{1 - \alpha_i} \tag{4}$$

where α_i is the proportion of amino acid substitutions that are adaptive (see Smith and Eyre-Walker [2002]). Expectations are summarized in table 1.

Table 1
Expected Numbers of Synonymous Polymorphisms ($\hat{P}_s$) and Substitutions ($\hat{D}_s$) and Numbers of Nonsynonymous Polymorphisms ($\hat{P}_n$) and Substitutions ($\hat{D}_n$)

	Polymorphisms	Substitutions
Synonymous	$\hat{P}_s = \Theta L$	$\hat{D}_s = \lambda L$
Nonsynonymous	$\hat{P}_s = \omega\Theta L$	$\hat{D}_n = (\omega\lambda + \eta)$ $= \omega\lambda L/(1-\alpha)$

NOTE.—L: length of the sequence (in codons); Θ: synonymous polymorphisms per codon; ω: nonsynonymous to synonymous diversity ratio ($= P_n/P_s$), which is used to measure the proportion of nonsynonymous polymorphisms that are neutral; λ: synonymous substitution rate per codon; η: number of adaptive amino acid substitutions per codon (providing that some assumptions hold); α: proportion of amino-acid substitutions that are adaptive (providing that some assumptions hold).

Parameter Estimation

Under the assumption of independence of sites (i.e., linkage equilibrium [Sawyer and Hartl 1992]), it is relatively easy to write the likelihood function because P_{ni}, P_{si}, D_{ni}, and D_{si} are Poisson distributed. The likelihood is

$$L = \prod_{i=1}^{n} H(\Theta_i L_i, P_{si}) H(\omega_i \Theta_i L_i, P_{ni}) H(\lambda_i L_i, D_{si}) \times H((\omega_i \lambda_i + \eta_i) L_i, D_{ni}) \tag{5}$$

where n is the number of loci analyzed and $H(\mu, x) = (e^{-\mu} \mu^x / x!)$ is the Poisson distribution. At, most this model has $4n$ parameters because there are four parameters per locus (Θ_i, λ_i, ω_i, and either η_i or α_i). However, we can reduce the number of parameters, by assuming either that a given parameter is constant across loci or that a parameter follows a probability density function. For example, we might proceed by assuming η or α are the same for all genes or that η is gamma distributed or α is beta distributed. When α is assumed to follow a distribution, the likelihood becomes

$$L = \prod_{i=1}^{n} \int_0^{\infty} PDF(\alpha) H(\hat{P}_{si}, P_{si}) H(\hat{P}_{ni}, P_{ni}) H(\hat{D}_{si}, D_{si}) \times H(\hat{D}_{ni}, D_{ni}) \partial\alpha \tag{6}$$

where $PDF(\alpha)$ is the probability distribution function of parameter α (see table 1 for the expectations of $\hat{P}_{si}$, $\hat{P}_{ni}$, $\hat{D}_{si}$ and $\hat{D}_{ni}$). Although the number of parameters is reduced, the numerical integration (that needs to be repeated at each step of the likelihood maximization [see below]) is computationally slow. To reduce computation times, we used the discrete distribution approximation method of Yang (Yang 1994; Yang et al. 2000). Briefly, K equiprobable classes of parameter x are assumed (p_0, p_1, ... , p_{K-1}, each with probability $1/K$), with the median value of x within each class to represent the distribution of x within that class. Let $CDF(x)$ be the cumulative distribution function of $PDF(x)$; that is, $CDF(x) = \int_0^x PDF(y) dy$. Let CDF^{-1} be the inverse cumulative distribution function. Then, for $j = 0$ to $K - 1$, we have $p_j = CDF^{-1}[(2j + 1)/(2K)]$. The following approximation is thus obtained:

$$\int_0^{\infty} PDF(x) F(x) \partial x \approx \frac{1}{K} \sum_{j=0}^{K-1} F\left[CDF^{-1}\left(\frac{2j+1}{2K}\right)\right] \tag{7}$$

where $F(x)$ is a function of x. The inverse cumulative gamma and beta distribution functions can be computed in Mathematica (Wolfram 1996) by using the functions "InverseGammaRegularized" and "InverseBetaRegularized," respectively.

The maximum-likelihood estimates are found by maximizing the log-likelihood, $\log(L)$. To prevent the maximization from the effect of initial parameter values, in a sometime bounded parameter space, and from the effect of local optima in this kind of parameter rich models, we chose to implement the Metropolis algorithm (Metropolis et al. 1954). A symmetrical uniform random change of length Δ_x is made to parameter x. If the new parameter is valid and increases $\log(L)$, it is accepted. If it is valid but decreases the likelihood by a factor $\rho < 1$, where ρ is the ratio of the likelihoods, then it is accepted with probability $\rho^{1/T}$. The size of random perturbations (Δ_x) is optimized by decreasing it slightly (i.e., $0.95\Delta_x$) if the change is rejected and by increasing it (i.e., $\Delta_x/0.95$) if a change is accepted. This procedure is applied to each of the parameters in turn. The parameter T, analogous to the temperature of the original algorithm, determines how close to the optimum (or optima) the random walk clusters. The global optimum is found by "simulating annealing," (Kirkpatrick, Gelatt, and Vecchi 1983) starting at some high T and gradually cooling to T = 0, following the procedure described in Barton (2000). We decided to use the Metropolis algorithm for convenience, but more direct maximization methods (such as direction set methods) appeared to give similar results when tested (not shown), although the choice of the starting points was sometimes problematic. The likelihood surface was investigated by incrementing the parameter of interest on the appropriate range around its maximum-likelihood value and maximizing all the other parameters. This procedure allowed us to obtain 2 units of $\log(L)$ confidence intervals. A Mathematica (Wolfram 1996) notebook performing the maximization is available from the authors.

The advantage of the likelihood framework is that it provides a natural way of comparing nested hypotheses (Mangel and Hilborn 1996; Barton 2000). Here, for instance, the first aim would be to compare the likelihood of a model without adaptive substitutions ($\eta = 0$ or $\alpha = 0$) with the likelihood of a model where adaptive substitutions can occur. One can then compare the likelihood of the model with a constant proportion of adaptive substitutions across genes ($\alpha_i = \alpha$ for all i) with the likelihood of a model that allows the proportion to vary across genes (beta

Table 2
Description, Number of Parameters and Log-Likelihood Values for Various Models

Model	Θ	λ	ω	η	α	NP	log(L)
1a	Θ (0.03)	λ (0.14)	1	—	—	2	−2037.5
1b	Θ (0.04)	λ (0.20)	ω (0.38)	—	—	3	−1592.1
1c	Θ (0.05)	λ (0.19)	ω (0.23)	η (0.037)	—	4	−1571.2
1d	Θ (0.05)	λ (0.19)	ω (0.23)	—	α (0.46)	4	−1571.2
2a	All Θ_i	All λ_i	All ω_i	—	—	$3n$ (105)	−331.2
2b	All Θ_i	All λ_i	All ω_i	η (0)	—	$3n$+1 (106)	−331.2
2c	All Θ_i	All λ_i	All ω_i	*Gamma*[a]	—	$3n$+2 (107)	−329.1
2d	**All Θ_i**	**All λ_i**	**All ω_i**	—	**α (0.26)**	**$3n$+1 (106)**	**−327.5**
2e	All Θ_i	All λ_i	All ω_i	—	*Beta*[b]	$3n$+2 (107)	−327.5
2f	All Θ_i	All λ_i	All ω_i	—	All α_i	$4n$ (140)	−302.9

NOTE.—Results obtained with the data set analyzed by Smith and Eyre-Walker (2002); Θ: synonymous polymorphisms per codon; λ: synonymous substitution rate per codon; ω: nonsynonymous to synonymous diversity ratio; η: number of adaptive amino acid substitutions per codon; α: proportion of amino acid substitutions that are adaptive; NP: number of parameters in the model; n: number of loci analyzed; log(L): log-likelihood. The best model according to the number of parameters and the likelihood is in bold type.

[a] See figure 2*A*.

[b] See figure 2*B*.

distribution or one α_i per locus). To compare different models, we have treated the likelihood itself as the criterion for inference. When hypotheses differ by several degrees of freedom, the appropriate model can be chosen using the Akaike information criterion (Mangel and Hilborn 1996) under which model 2 is preferred to model 1 if $\log(L_2) - 2\nu > \log(L_1)$, where ν is the number of additional parameters used in model 2 relative to model 1. It should be emphasised that comparative tests assume that recombination is a stronger force than mutation. so the population approximates a state of quasilinkage equilibrium; if this is not the case, then the confidence intervals are underestimated and the tests become liberal.

Results

Maximum-Likelihood Estimation

In this section, we describe the results obtained with the data set analyzed by Smith and Eyre-Walker (2002), but qualitatively similar results were obtained with other data sets. To begin, let us consider the simplest models where each parameter is constant for all genes (model 1a to model 1d in table 2). A model with only two parameters (model 1a), Θ and λ, is massively improved (approximately 445 units of log(L)) by adding the parameter ω (model 1b). This is because the vast majority of nonsynonymous mutations are deleterious. Adding either parameter η (model 1c) or α (model 1d) substantially improves the model (approximately 20 units of log(L)) and leads to an estimation of η approximately 0.04 adaptive substitutions per codon or α approximately 46%. Both of these estimates are significantly different from zero.

We have so far assumed that all parameters are constant across genes, which seems very unlikely. Let us first consider the three parameters Θ, λ, and ω. First we assumed that Θ, λ, and ω were gamma distributed. The estimated distributions of each of the three parameters are superimposed on the observed distributions in figure 1. To obtain the observed distributions, we used the level of synonymous variation and substitutions to obtain point estimates of Θ, λ, and ω using the following equations:

$$\Theta = \frac{P_s}{L}, \qquad \lambda = \frac{D_s}{L}, \qquad \omega = \frac{P_n}{P_s} \tag{8}$$

The fit appeared to be quite good (fig. 1), however a per-locus parameterized model (in which each of the $3n$ different parameters Θ_i, λ_i, and ω_i are simultaneously maximized) was significantly better supported. In addition, even if the number of parameters being maximized is increased, the maximization for the per-locus model is much more rapid, taking on average minutes rather than days or weeks, as the maximization takes for a model involving several gamma or beta distributions. This per-locus parameterization for Θ, λ, and ω was, therefore, the one used hereafter.

We can now reconsider our parameters of interest, beginning with η, the number of adaptive substitutions per codon. If we add to a model without adaptive substitutions (model 2a), a single η equal for all genes (model 2b), the likelihood remains unchanged and an estimate of roughly no adaptive substitutions per codon is obtained (table 2). However, if we consider η to be gamma distributed (model 2c), the Akaike information criterion is just crossed and a leptokurtic (i.e., L-shaped) distribution is obtained (fig. 2*A*).

Interestingly, the proportion of adaptive substitutions, α, does not behave in a similar manner. Adding the same proportion for all genes (model 2d) significantly increases the log-likelihood (approximately 4 units of log(L)). The estimated proportion is approximately 25%. However, allowing α to vary by a beta distribution does not improve the model (model 2e), because the resulting variance across genes is very low (fig. 2*B*). The Akaike information criterion is also not crossed when a per-locus parameterization is used for α (model 2f). The log-likelihood for a model with a constant α across genes (model 2d), which is the best model according to its number of parameters and its likelihood, is plotted as a function of α in figure 3. The 2 units of Log(L) confidence interval is estimated to be [0.08, 0.41].

The apparent constancy of α across genes implies a linear relationship between the number of adaptive

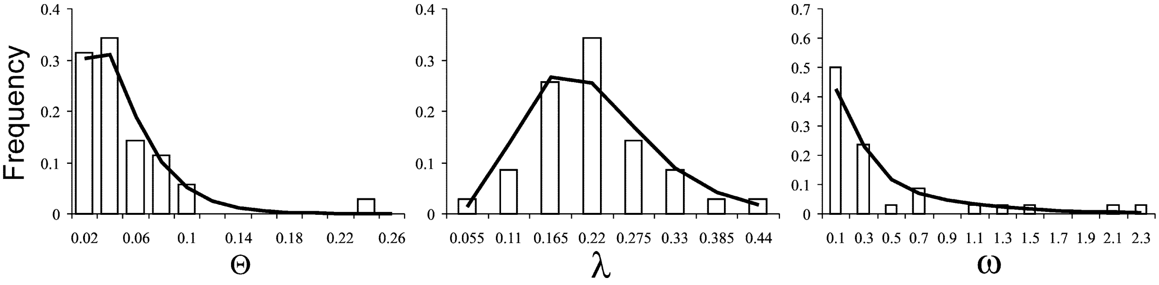

FIG. 1.—Observed (histograms) and estimated (lines) distributions of the three parameters, Θ (synonymous diversity per codon), λ (synonymous substitution rate per codon), and ω (nonsynonymous to synonymous diversity ratio). Results obtained with the data set analyzed by Smith and Eyre-Walker (2002).

nonsynonymous substitutions and the number of nonsynonymous substitutions. To obtain a visual representation of this relationship, we have implemented a random walk with parameter T (temperature of the simulating annealing) set up to 1, allowing us to infer the marginal distribution of each η_i weighting the other parameters by their likelihood. The procedure amounts to Bayesian inference with a uniform prior (Barton 2000). The correlation between the mean of the posterior distribution and the number of nonsynonymous substitutions per codon, d_n, is illustrated in figure 4*A*. The absence of a correlation between the proportion of adaptive substitutions, α, and the nonsynonymous substitution rate per codon, was obtained in a similar way and is illustrated in figure 4*B*.

Comparative Analysis

We have estimated the rate of adaptive amino acid substitutions for various other *Drosophila* data sets using either polymorphism in *D. melanogaster* or in *D. simulans* (table 3). As stated above, we used the per-locus parameterization for Θ, λ, and ω because it gives the highest likelihoods and is much faster to evaluate. We compared a model without adaptive substitutions ($\alpha = 0$, model 2a in table 2) with a model where the rate of adaptive substitutions is constant across genes ($\alpha_i = \alpha$ for all i, model 2d in table 2). As in the previous section, we never obtained a significant increase of the likelihood using a model that allows the proportion of adaptive changes to vary across genes (beta distribution or one α_i per locus).

Using polymorphism data in *D. simulans*, we always obtained significant estimates of α (table 3). The α estimates were higher when divergence data included the *D. simulans* lineage, were intermediate when divergence data included the *D. melanogaster* lineage, and were smallest when divergence data included the *D. yakuba* lineage. However, confidence intervals were large. Removing singletons from the polymorphism data had virtually no effect on the estimates because the ratio P_n/P_s is invariant among frequency classes in *D. simulans*. The number of genes used does not affect the estimate of α greatly, which is in agreement with the apparent constancy of α across genes.

Using polymorphism data in *D. melanogaster*, the estimate of α was different among data sets and sometimes not significantly different from zero (table 3). Most importantly, the procedure of removing singletons had a dramatic effect. When all the polymorphism data were used, there was no evidence of adaptive evolution, except when the *D. simulans* lineage was included in the divergence. However, when singletons were removed from the data set, a significant estimate was restored whatever the divergence data set. This is because the ratio P_n/P_s does vary among frequency classes in *D. melanogaster* (see Akashi [1996] and Fay, Wycoff, and Wu [2002]). The estimates of α obtained when removing singletons from the polymorphism data were similar to those obtained using the polymorphism data from *D. simulans*.

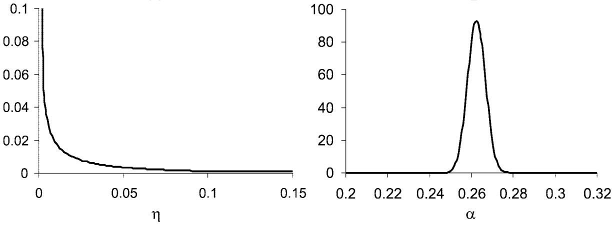

FIG. 2.—(*A*) Maximum-likelihood gamma-distribution for parameter η, the number of adaptive amino acid substitutions per codon. (*B*) Maximum-likelihood beta-distribution for parameter α, the proportion of adaptive amino acid substitutions. Results obtained with the data set analyzed by Smith and Eyre-Walker (2002).

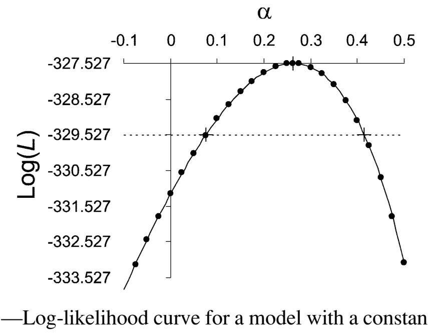

FIG. 3.—Log-likelihood curve for a model with a constant proportion of adaptive substitutions (α) across genes (model 2d in table 2). Dots are maximum log-likelihood values for the α-values considered and the solid line is a polynomial adjustment. The dotted line is 2 units of log(*L*) below the maximum log-likelihood. Crosses highlight the lower and upper limits of the 2 units of log(*L*) confidence interval and the maximum-likelihood estimate of α. Results obtained with the data set analyzed by Smith and Eyre-Walker (2002).

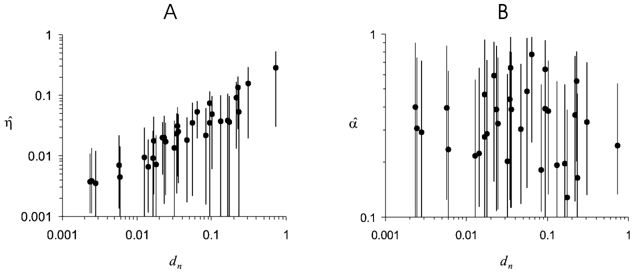

FIG. 4.—(*A*) Relationship between the estimate of the number of adaptive amino acid substitutions per codon, $\hat{\eta}$, and the rate of amino acid substitutions per codon, d_n. (*B*) Relationship between the estimate of the proportion of adaptive amino acid substitutions, $\hat{\alpha}$, and d_n. Axes are on a log scale. Results obtained with the data set analyzed by Smith and Eyre-Walker (2002).

The question arises whether α significantly varies among lineages. To test for such variation, we used the 22 genes for which we had polymorphism data in both *D. melanogaster* and *D. simulans* and divergence data in the three lineages: *D. melanogaster*, *D. simulans*, and *D. yakuba*. Considering the results obtained using polymorphism data in *D. melanogaster*, singletons were removed from polymorphism data in both species. The likelihood of different nested models was estimated: we first assumed that α was constant in the three lineages, then allowed α to be different in one of the three lineages in turn, and finally allowed α to vary among the three lineages. Note, however, that in each model, α remains constant across genes within a lineage. Results are presented in table 4 and imply a significant higher rate of adaptive evolution in the *simulans* lineage ($\hat{\alpha}_{sim} \sim 67\%$) but no significant differences between the *melanogaster* and the *yakuba* lineages ($\hat{\alpha}_{mel} = \hat{\alpha}_{yak} \sim 40\%$).

Testing the Effect of Codon Usage Bias and Linkage Disequilibrium

As expected, from the apparent constancy of a across genes, α did not differ significantly between the two halves of the data set when the data set was split according to codon usage bias or linkage disequilibrium (table 5).

All else being equal, weak selection for synonymous codon use is expected to reduce the ratio D_s/P_s and therefore inflate the estimate of α (see equation 1). However, if the strength of selection acting on synonymous mutations (γ_s) is smaller than that acting on weakly selected nonsynonymous mutations (γ_n), the rate of adaptive amino acid substitution is expected to be underestimated (Charlesworth 1994; Eyre-Walker 2002). Accordingly, α was lower for biased genes than for unbiased genes, although the difference was not significant. The available evidence suggests that it is likely that $|\gamma_s| \leq |\gamma_n|$ because nonsynonymous polymorphisms tend to segregate at either the same frequency (as in *D. simulans*) or lower frequencies (as in *D. melanogaster*) than synonymous polymorphisms. However, to further investigate the effect of selection on synonymous codon use, we have undertaken an analysis in which mutations in introns are used as the neutral reference instead of synonymous mutations. Results are presented in table 6. Estimates of α are virtually unchanged whether intron mutations or synonymous mutations are used. Taken together, these results suggest that the effect of synonymous selection on our estimates is very weak, even for highly biased genes.

The amount of linkage disequilibrium actually was not expected to bias the estimate of adaptive evolution but rather to bias tests and confidence intervals. However, we obtained very similar estimates and confidence intervals when we restricted the analysis to genes with strong linkage disequilibria (CI = 33%) or to genes with slight linkage disequilibria (CI = 40%).

Discussion

We have used a simple likelihood method that allowed us to combine data across genes to estimate the average rate of adaptive substitution and to compare nested hypotheses. Application of the method to data from *Drosophila* suggests that a substantial fraction of amino acid substitutions have been driven by positive adaptive

Table 3
Estimates of α, the Proportion of Amino Acid Substitutions That Are Adaptive, in Various *Drosophila* Data Sets

Species for Polymorphism	Singletons	Species for Divergence	Number of Genes	$\Delta\log(L)$	$\hat{\alpha}$ [2 Units of log(*L*) CI]	$\hat{\gamma}$
D. simulans	Included	*D. simulans–D. yakuba*	35	**3.7**	**0.26** [0.08, 0.41]	0.48
D. simulans	Included	*D. simulans–D. melanogaster*	75	**13.9**	**0.43** [0.30, 0.55]	0.98
D. simulans	Included	*D. simulans–D. melanogaster*	44	**18.6**	**0.55** [0.38, 0.66]	1.50
D. simulans	Excluded	*D. simulans–D. melanogaster*	44	**9.9**	**0.53** [0.37, 0.65]	1.40
D. simulans	Included	*D. simulans* lineage	22	**9.8**	**0.65** [0.43, 0.78]	2.16
D. simulans	Excluded	*D. simulans* lineage	22	**7.0**	**0.63** [0.39, 0.78]	2.02
D. simulans	Included	*D. melanogaster* lineage	22	**8.4**	**0.56** [0.31, 0.68]	1.56
D. melanogaster	Included	*D. melanogaster–D. yakuba*	25	0.1	0.06 [−0.15, 0.25]	0.09
D. melanogaster	Excluded	*D. melanogaster–D. yakuba*	25	**2.5**	**0.31** [0.08, 0.47]	0.61
D. melanogaster	Included	*D. melanogaster–D. simulans*	57	**2.2**	**0.22** [0.02, 0.44]	0.38
D. melanogaster	Included	*D. melanogaster–D. simulans*	44	**2.2**	**0.26** [0.02, 0.44]	0.48
D. melanogaster	Excluded	*D. melanogaster–D. simulans*	44	**4.2**	**0.45** [0.15, 0.62]	1.06
D. melanogaster	Included	*D. melanogaster* lineage	22	0.2	−0.13 [−0.35, 0.27]	−0.18
D. melanogaster	Excluded	*D. melanogaster* lineage	22	0.1	0.10 [−0.31, 0.49]	0.16
D. melanogaster	Included	*D. simulans* lineage	22	**6.7**	**0.64** [0.38, 0.78]	2.10

NOTE.—$\Delta\log(L)$: difference between the log-likelihood of a model without adaptive substitution ($\alpha = 0$, model 2a in table 2) and the log-likelihood of a model with a constant rate of adaptive substitution across genes ($\alpha_i = \alpha$ for all i, model 2c in table 2); significant values (i.e., $\Delta\log(L) > 2$) are in bold. $\hat{\gamma}$: estimate of the strength of selection, $2N_es$, under the model of Sawyer and Hartl (1992).

Table 4
Variation of the Estimate of α Among Lineages

Model	NP	Log(L)	$\hat{\alpha}_{mel}$	$\hat{\alpha}_{sim}$	$\hat{\alpha}_{yak}$
$\alpha_{mel} = \alpha_{sim} = \alpha_{yak}$	$6n + 1$ (133)	−416.28	0.43	0.43	0.43
$\alpha_{mel} = \alpha_{sim}, \alpha_{yak}$	$6n + 2$ (134)	−409.00	0.58	0.58	0.37
$\alpha_{mel} = \alpha_{yak}, \alpha_{sim}$	**$6n + 2$ (134)**	**−407.78**	**0.39**	**0.67**	**0.39**
$\alpha_{sim} = \alpha_{yak}, \alpha_{mel}$	$6n + 2$ (134)	−415.71	0.49	0.42	0.42
$\alpha_{mel}, \alpha_{sim}, \alpha_{yak}$	$6n + 3$ (135)	−406.18	0.50	0.68	0.37

NOTE.—Results obtained with 22 genes with polymorphism data in *D. melanogaster* and *D. simulans* (singletons are excluded) and divergence data in the three lineages *D. melanogaster*, *D. simulans*, and *D. yakuba*. NP: number of parameters in the model; α_{mel}: estimate of α in the *melanogaster* lineage; α_{sim}: estimate of α in the *simulans* lineage; α_{yak}: estimate of α in the *yakuba* lineage. The best model according to the number of parameters and the likelihood is in bold type.

evolution in these species. Surprisingly, the proportion seems to be remarkably constant across genes. However, estimates were sometimes different among data sets, and we have made a number of simplifying assumptions.

Underlying Assumptions

First, we have assumed that synonymous mutations are neutral, although there is evidence that selection for codon usage is acting upon synonymous mutations in some organisms, including *Drosophila* (Shields et al. 1988; Akashi 1995). Evidence from polymorphism data suggests that weak selection for codon usage seems to be relaxed in *D. melanogaster* (Akashi 1996) but is currently active in *D. simulans* (Akashi and Schaeffer 1997; Kliman 1999; Begun 2001). However, we found no evidence that our estimates of α are substantially biased by synonymous selection. The estimates remained unchanged (1) when we restricted the analysis to genes with low codon bias (table 5) or (2) when mutations in the introns were used as the neutral reference instead of synonymous mutations (table 6).

Second, we have assumed independence among sites to compute the likelihood function. This assumption is unlikely to strongly bias estimates of α, but we are aware that the violation of this assumption would bias our tests and confidence intervals. We found no simple solution to this problem. However, linkage disequilibria did not appear to affect confidence intervals much, because we obtained very similar results whether we restricted our analysis to genes with high or low linkage disequilibria.

Third, and most importantly, we have assumed that nonsynonymous mutations segregating within a species are neutral. However, it is known that slightly deleterious mutations can segregate at a nonnegligible frequency (Crow and Kimura 1970); in humans and *D. melanogaster*, the ratio P_n/P_s is higher in rare than in common polymorphisms, suggesting that a fraction of nonsynonymous mutations are slightly deleterious (Akashi 1996; Cargill et al. 1999; Fay and Wu 2001). This will make the McDonald-Kreitman approach conservative if population sizes have been roughly constant or have contracted. However, artifactual evidence of adaptive evolution can be produced if the current effective population size is larger than the long-term effective population size (McDonald and Kreitman 1991; Eyre-Walker 2002; Fay, Wycoff, and Wu 2002). This is because slightly deleterious mutations may have been fixed in the past that can no longer segregate as polymorphisms. The difference in effective population size does not have to be very great to generate artifactual evidence of adaptive evolution when synonymous mutations are neutral (Eyre-Walker 2002). However, selection on synonymous codon use restricts the conditions under which artifactual evidence of adaptive evolution is produced (Eyre-Walker 2002).

The rate of adaptive amino acid substitution does not significantly depart from constancy across the genome. Fay, Wycoff, and Wu (2002) previously argued that the evidence of adaptive evolution in *D. melanogaster* was not a consequence of slightly deleterious mutations and an expansion in population size, because the evidence of adaptive evolution in their data seemed to be restricted to the most rapidly evolving genes. They argued that an increase in effective population size would tend to affect all genes in a similar fashion. Our results contradict their findings. Although rapidly evolving genes undergo more adaptive substitution, the proportion of substitutions that appear to be adaptive seems constant across genes. The constancy is supported by the fact that a model that allows the proportion to vary between genes is not significantly better than a model with a constant proportion. Furthermore, if we model variation in α using a beta distribution, the estimated variance is small (fig. 2*B*). This assertion is also compatible with the results obtained by Bustamante et al. (2002), in which there is no trend for genes with a higher estimated selection intensity ($\gamma = 2N_es$) to have a higher rate of nonsynonymous substitutions.

However, the constancy of α does not tell us whether the evidence for adaptive evolution is artifactual. If all genes had the same distribution of fitness effects (for deleterious mutations), then we might expect an increase in population size to have a similar proportional effect on all genes.

Table 5
The Effect of Codon Usage Bias and Linkage Disequilibrium on the Estimate of α

Criteria Used to Split the Data Set	Number of Genes	Average of the Criteria (*Fop* or *LD*)	$\hat{\alpha}$ [2 Units of log(L) CI]
Low *Fop*	37	0.47	0.48 [0.34, 0.60]
High *Fop*	37	0.68	0.33 [−0.05, 0.59]
Low *LD*	15	0.31	0.50 [0.31, 0.64]
High *LD*	15	0.59	0.45 [0.22, 0.62]

NOTE.—Results obtained with the largest data set composed of 75 genes with polymorphism data in *D. simulans* and divergence between *D. simulans* and *D. melanogaster*. *Fop*: Frequency of optimal codons; *LD*: linkage disequilibrium (average of r^2).

Table 6
Comparison of α-Estimates When Synonymous or Intron Mutations Are Used As the Neutral Reference

Neutral Reference	Species for Polymorphism	Singletons	Number of Genes	$\Delta\log(L)$	$\hat{\alpha}$ [2 Units of $\log(L)$ CI]
Synonymous	*D. simulans*	Included	34	15.9	0.51 [0.37, 0.62]
Introns	*D. simulans*	Included	34	6.4	0.45 [0.23, 0.61]
Synonymous	*D. melanogaster*	Included	38	0.4	0.11 [−0.16, 0.33]
Introns	*D. melanogaster*	Included	38	0.6	0.16 [−0.17, 0.39]
Synonymous	*D. melanogaster*	Excluded	38	3.0	0.35 [0.08, 0.54]
Introns	*D. melanogaster*	Excluded	38	0.9	0.25 [−0.15, 0.52]

NOTE.—Results obtained with divergence data between *D. melanogaster* and *D. simulans*. $\Delta\log(L)$: difference between the log-likelihood of a model without adaptive substitution ($\alpha = 0$) and the log-likelihood of a model with a constant rate of adaptive substitution across genes ($\alpha_i = \alpha$ for all i).

However, there is a priori no reason to expect the distribution to be similar for genes with different functions and rates of evolution. Still, it is also puzzling why the proportion should be constant if the adaptive evolution is genuine.

Is the evidence of adaptive evolution artifactual? Because there is evidence that some nonsynonymous mutations are slightly deleterious, this question amounts to a question of whether the current effective population sizes of the *Drosophila* species considered here are larger than they have been on average in the past. This is a difficult question to answer, not only because current *Drosophila* populations show complex patterns that are not consistent with any simple model (Moriyama and Powell 1996; Powell and Moriyama 1997; Andolfatto and Przeworski 2000; Begun 2001; Wall, Andolfatto and Przeworski 2002) but also because we are attempting to look back at something that is very difficult to measure in the past.

D. simulans is thought to have had a fairly stable population size (Li, Satta, and Takahata 1999; Takahata and Satta 2002). Although, synonymous codon bias is declining in *D. simulans* (Begun 2001; McVean and Vieira 2001), selection has continued to operate since the split with *D. melanogaster* (Akashi 1995; McVean and Vieira 2001) and is detectable in current polymorphisms (Akashi and Schaeffer 1997; Kliman 1999; Begun 2001). These observations suggest that the effective population size has not changed greatly or has decreased, a view corroborated by two estimates of $N_e s$ on synonymous codons in *D. simulans*. These two estimates come from a comparison of substitution and polymorphism data (Akashi 1995) and from the allelic frequencies of synonymous polymorphisms (Akashi and Schaeffer 1997). The two estimates are very similar (Akashi and Schaeffer 1997). Furthermore, the presence of selection on synonymous codon use greatly restricts the conditions under which artifactual evidence of adaptive evolution can be produced (Eyre-Walker 2002). The evidence, therefore, suggests that the current effective population size of *D. simulans* is similar to the long-term effective population size experienced by *D. simulans* since it split from *D. melanogaster* and that estimates of adaptive evolution in this lineage are likely to be accurate.

In contrast to *D. simulans*, several lines of evidence suggest that *D. melanogaster* has decreased in effective population size: (1) there has been a sharp decrease in synonymous codon bias (Akashi 1996), (2) there is no evidence of current selection on synonymous codon use (Akashi and Schaeffer 1997; McVean and Vieira 2001), and (3) the synonymous diversity is lower in *D. melanogaster* than in *D. simulans* or *D. pseudoobscura* (Moriyama and Powell 1996). It has also previously been suggested that the nonsynonymous relative to the synonymous substitution rate has increased in the *D. melanogaster* lineage (Akashi 1996, Eyre-Walker et al. 2002). This has been interpreted as being the consequence of the fixation of slightly deleterious amino acid mutations in this lineage. However, our reanalysis of this pattern using only genes for which we had polymorphism data in both *D. melanogaster* and *D. simulans* suggests that the increase in the nonsynonymous substitution rate was an artifact of counting some polymorphic sites in the divergence when only one sequence is used to compute the substitution rate. In other words, slightly deleterious mutations are segregating in the polymorphism but have not fixed (or, at least, have not fixed yet).

We, therefore, suspect that the use of polymorphism data from *D. melanogaster* is likely to lead to an underestimation of the adaptive substitution rate, unless a correction for slightly deleterious mutations is found. The removal of low frequency polymorphisms (Fay, Wycoff, and Wu 2002) may provide such a correction. Accordingly, we obtained very similar estimates of adaptive evolution when singletons were removed from the polymorphism data in *D. melanogaster* as we did using all the polymorphism data in *D. simulans* (table 3).

However, we obtained different results depending on the lineage used for the divergence. Estimates from the *simulans* lineage are significantly higher than from the other two lineages, which do not differ significantly. There are at least two explanations for this. First, *D. simulans* may have gone through more adaptive evolution than the other two species. This is a rather ad hoc explanation. Second, the difference may be caused by the fact that *D. simulans* has not reached, as Lewontin (2002) describes it, a “stochastic steady sate.” Substitutions become fixed during the period between the time of speciation and the time when the alleles coalesce. Because the effective population size of *D. simulans* is higher than that of *D. melanogaster*, the coalescence time is on average longer in

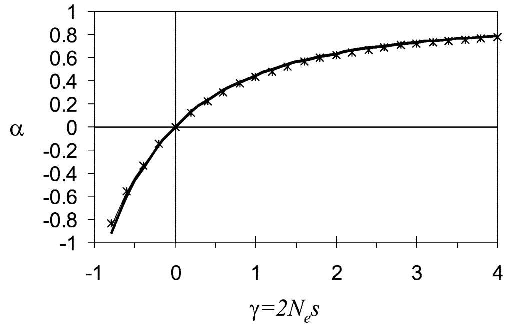

FIG. 5.—The proportion of amino acid substitutions exceeding the expectation of an equal nonsynonymous to synonymous ratio in polymorphism and divergence, α, as a function of $\gamma = 2N_es$ (where N_e is the effective population size and s the selection coefficient). Sawyer and Hartl's (1992) sampling formulae were used with a sample size of five (crosses) and 10 (line) sequences (the two curves are roughly superposed).

the former and there is, therefore, less time for fixation. Furthermore, the time to coalescence forms a substantial fraction of the total divergence along the *D. simulans* lineage. Therefore, neutral mutations will have had less time to fix in the *simulans* lineage. Accordingly, the number of synonymous substitutions in the *simulans* lineage is on average approximately two-thirds that in the *melanogaster* lineage. However, the short time available for fixation will not be such a problem for advantageous mutations, because they spread much more rapidly through a population than neutral mutations.

So it is possible that the estimate of adaptive evolution obtained using the *simulans* lineage is an overestimate. However, the time for substitutions between either *D. simulans* or *D. melanogaster* and *D. yakuba* should be sufficiently long for stochastic steady state to have been established for both adaptive and neutral mutations. It, therefore, appears preferable to use the divergence data that includes *D. yakuba*. In conclusion, our best estimate is that approximately 25% ± 20% of amino acid substitutions were driven by positive selection in the divergence between *D. simulans* and *D. yakuba*.

Alternative Models and the Fitness Effect of Adaptive Substitutions

We have assumed so far that nonsynonymous mutations are strongly deleterious, neutral, or strongly advantageous. Under this simple distribution of fitness effects, we were able to estimate the rate of adaptive evolution, but there is no way within our method to separately estimate the fitness effects and number of adaptive substitutions. There are two solutions to this problem. The first is to assume that all mutations that contribute to polymorphism and substitution have the same strength of selection acting upon them and then to estimate the strength of this selection (Sawyer and Hartl 1992; Bustamante et al. 2002). Under the Sawyer-Hartl model, nonsynonymous mutations are either strongly deleterious or weakly selected, with all mutations in the weakly selected class being subject to the same strength of selection. In this model, it is further assumed that the effective population size, N_e, is constant through time, that the actual population size, N, is of the same order of magnitude as the effective size (i.e., $N_e/N \approx 1$) and that the strength of selection is sufficiently weak for diffusion approximations to be used. Under this viewpoint, polymorphisms and substitutions are coupled and share the same constant intensity of selection (N_es), which is the parameter of interest in this case. Bustamante et al. (2002) have developed a method to estimate $\gamma = 2N_es$ from multilocus data. These authors have applied their method to a *Drosophila* data set with polymorphism data from *D. melanogaster* and divergence data between *D. melanogaster* and *D. simulans*. They obtained an estimate for γ of approximately 1.5.

Because the α parameter is a combination of D_n, D_s, P_n, and P_s (see equation 1), the sampling formula of Sawyer and Hartl (1992) can be used to convert our estimates of α into an estimate of γ for a single gene (see Akashi [1995]), although this involves modifying the underlying assumptions about the distribution of fitness effects in the model quite radically. Furthermore, it turns out to be easy get an average estimate of γ across genes, because the relationship between α and γ is largely independent of sample size (fig. 5). Estimates of γ are given in table 3. Our estimates are of the same order of magnitude as those of Bustamante et al. (2002), although our estimates from data sets comparable with theirs appeared to be slightly lower: γ is approximately 0.5 for *D. melanogaster–D. simulans* using polymorphism data from *D. melanogaster*. The slight discrepancy may be caused by the fact that Bustamante et al. (2002) assumed that all the genes share the same divergence time, scaled in N_e generations, whereas we have relaxed this constraint.

Assuming a single strength of selection is unlikely to be very realistic. However, additional information may help us to further investigate the fitness effects of adaptive substitutions. Stephan and collaborators (Wiehe and Stephan 1993; Stephan 1995; Kim and Stephan 2000, 2003) have developed a series of models of recurrent selective sweeps to explain the positive correlation between neutral diversity and local recombination rates in *D. melanogaster*. For the model to fit the data, the intensity of directional selection, $\gamma\nu$ (where $\gamma = 2N_es$ and ν is the rate of selected substitution per nucleotide per generation), has to be somewhere between 10^{-8} and 10^{-7} (Stephan 1995; Andolfatto 2001). Our estimate of α at approximately 25% transforms into a rate of selected substitution per nucleotide per generation for ν of approximately 10^{-11}. Our estimates of ν appears to be in rough accordance with the estimate for ν of approximately $6 \times 10^{-6}\ N^{-1}$ obtained by Stephan and Kim (2002) from the results of Perlitz and Stephan (1997), if we assume that N is approximately 10^6 in *Drosophila*. Therefore, the recurrent selective sweep model accounts for the diversity/recombination correlation if $10^3 < \gamma < 10^4$, three or four orders of magnitude higher than the estimate of Bustamante et al. (2002). The discrepancy is not surprising, because it seems very likely that mutations of different types will be responsible for polymorphism and divergence; the former is probably largely made up of neutral and slightly deleterious mutations, whereas a major contributor to the latter may be strongly advantageous mutations that contribute little to polymorphism.

Conclusion

In summary, evidence suggests that the estimate of adaptive evolution is not artifactual and that a substantial fraction of amino acid substitutions have been driven by positive adaptive evolution. Our best estimate of this fraction is 25% ± 20% (or one adaptive substitution approximately every 800 ± 350 generations). Such a nonnegligible genome-wide value is sufficient to have important consequences for evolutionary biology.

Appendix

The Effect of Slightly Advantageous Mutations on the Estimation of α

In this appendix, we investigate the bias caused by neglecting slightly advantageous mutations in the polymorphism. Assuming a standard Wright-Fisher model of evolution, the probability that we will observe a codominant mutation with selective advantage s in a sample of n sequences is

$$\prod(\gamma, \theta) = \theta \int_0^1 (1 - x^n - (1 - x)^n) \times \frac{(1 - e^{-2\gamma(1-x)})}{(1 - e^{-2\gamma})x(1 - x)} \partial x \qquad \text{(A1)}$$

where $\gamma = 2N_e s$, $\theta = 4N_e u$, and u is the mutation rate per generation (Sawyer and Hartl 1992; Eyre-Walker 2002). Kimura (1983) has shown that the rate of substitution per generation of such a mutation is

$$Q(\gamma, u) = u\frac{2\gamma}{1 - e^{-2\gamma}} \qquad \text{(A2)}$$

Let us assume that all synonymous mutations are neutral and that a proportion α^* of nonsynonymous mutations are slightly advantageous, all being subject to the same intensity of selection γ, and that a proportion $1 - \alpha^*$ are neutral. In our method, the proportion of nonsynonymous mutations that are advantageous would be estimated as

$$\hat{\alpha} = 1 - \frac{[\alpha^* \prod(\gamma, \theta) + (1 - \alpha^*) \prod(0, \theta)]u}{\prod(0, \theta)[\alpha^* Q(\gamma, u) + (1 - \alpha^*)u]} \qquad \text{(A3)}$$

an expression that is solely a function of α^* and γ because the mutation parameters, θ and u, cancel out. We define the bias caused by neglecting slightly advantageous mutations in the polymorphism in our method as being $\varepsilon = (\hat{\alpha} - \alpha^*)/\alpha^*$ (which is a function of γ only). The bias ε is plotted against γ in figure A1. The bias is always negative (i.e., α^* is underestimated) and quickly decreases when γ increases. For instance, if all the adaptive mutations have a fitness effect of $\gamma = 10$, the proportion of adaptive substitution is only underestimated by approximately 10% of its actual value. In the present model, we consider an extreme situation because all the adaptive substitutions are weakly selected. However, the bias will be smaller as soon as the fraction of strongly selected substitutions increases.

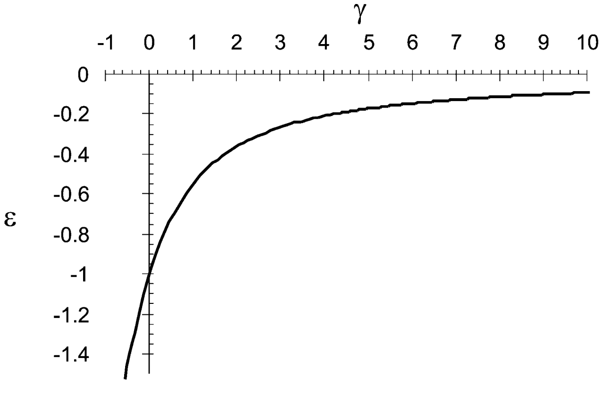

FIG. A1.—Relationship between the bias caused by neglecting slightly selected mutations in the polymorphism ($\varepsilon = (\hat{\alpha} - \alpha^*)/\alpha^*$, where $\hat{\alpha}$ is the estimate and α^* is the true value) and the intensity of selection ($\gamma = 2N_e s$).

Acknowledgment

We are very grateful to Nick Smith for sharing his data sets and to Gwenael Piganeau and two anonymous referees for helpful comments on the manuscript. The authors are supported by the BBSRC and the Royal Society.

Literature Cited

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

———. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144**:1297–1307.

———. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151**:221–238.

Akashi, H., and S. W. Schaeffer. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. Genetics **146**:295–307.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. **11**:635–641.

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

Barton, N. H. 2000. Estimating multilocus linkage disequilibria. Heredity **84**:373–389.

Begun, D. 2001. The frequency distribution of nucleotide variation in *Drosophila simulans*. Mol. Biol. Evol. **18**:1343–1352.

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

Brookfield, J. J. Y., and P. M. Sharp. 1994. Neutralism and selectionism face up to DNA data. Trends Genet. **10**:109–111.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in Arabidopsis. Nature **416**:531–534.

Cargill, M., D. Altshuler, J. Irel et al. (17 co-authors). 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22**:231–238.

Chakraborty, R., P. A. Fuerst, and M. Nei. 1978. Statistical studies on protein polymorphism in natural population. II. Gene differentiation between populations. Genetics **88**: 367–390.

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63**:213–227.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper and Row, New York.

Eyre-Walker, A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics **162**:2017–2024.

Eyre-Walker, A., P. D. Keightley, N. G. C. Smith, and D. Gaffney. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol. Biol. Evol. **19**:2142–2149.

Fay, J. C., G. J. Wycoff, and C.-I. Wu. 2001. Positive and negative selection on the human genome. Genetics **158**:1227–1234.

———. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415**:1024–1026.

Fay, J. C., and C. I. Wu. 2001. The neutral theory in the genomic era. Curr. Opin. Genet. Dev. **11**:642–6.

Goldman, N., and Z. Yang. 1994. A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**:226–231.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

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

———. 2003. Selective sweeps in the presence of interference among partially linked loci. Genetics **164**:389–398.

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

Kimura, M., and T. Ohta. 1971. Protein polymorphisn as a phase of molecular evolution. Nature **229**:467–469.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. Science **220**:671–680.

Kliman, R. 1999. Recent selection on synonymous codon usage in *Drosophila*. J. Mol. Evol. **49**:343–351.

Kreitman, M., and H. Akashi. 1995. Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26**:403–422.

Lewontin, R. C. 2002. Directions in evolutionary biology. Annu. Rev. Genet. **36**:1–18.

Li, Y. J., Y. Satta, and N. Takahata. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. **74**:117–127.

Mangel, M., and R. Hilborn. 1996. The ecological detective. Princeton, NJ.

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

McVean, G., and J. Vieira. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics **157**:245–257.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1954. Equation of state calculations by fast computing machines. J. Chem. Phys. **21**:1087–1095.

Moriyama, E. N., and J. R. Powell. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**:261–277.

Nachman, M. W., V. L. Bauer, S. L. Crowell, and C. F. Aquadro. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics **150**:1133–1141.

Perlitz, M., and W. Stephan. 1997. The mean and variance of the number of segregating sites since the last hitchhiking event. J. Math. Biol. **36**:1–23.

Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**:7784–7790.

Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174–175.

Sawyer, S. A., and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. Genetics **132**:1161–1176.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. "Silent" sites in *Drosophila* are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

Skibinski, D. O. F., and R. D. Ward. 1982. Correlations between heterozygosity and evolutionary rate of proteins. Nature **298**:490–492.

Smith, N. G. C., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. Nature **415**:1022–1024.

Stephan, W. 1995. An improved method for estimating the rate of fixation of favourable mutations based on DNA polymorphism data. Mol. Biol. Evol. **12**:959–962.

Stephan, W., and Y. Kim. 2002. Recent applications of diffusion theory to population genetics. Pp. 72–93 *in* M. Slatkin and M. Veuille, eds. Modern developments in theoretical population genetics, the legacy of Gustave Malecot. Oxford University Press, Oxford, UK.

Takahata, N., and Y. Satta. 2002. Pre-speciation coalescence and the effective size of ancestral populations. Pp. 52–71 *in* M. Slatkin and M. Veuille, eds. Modern developments in theoretical population genetics, the legacy of Gustave Malecot. Oxford University Press, Oxford, UK.

Wall, J. D., P. Andolfatto, and M. Przeworski. 2002. Testing models of selection and demography in *Drosophila simulans*. Genetics **162**:203–216.

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

Wolfram, S. 1996. The mathematica book. 3rd Edition. Cambridge University Press, Cambridge, UK.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1999. Phylogenetic analusis by maximum likelihood (PAML). Version 2. University College London, England.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Wolfgang Stephan, Associate Editor

Accepted March 4, 2004