Inferring Parameters of Mutation, Selection and Demography From Patterns of Synonymous Site Evolution in Drosophila

Gilean A. T. McVean and Jorge Vieira

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received June 19, 2000
Accepted for publication September 27, 2000

ABSTRACT

Selection acting on codon usage can cause patterns of synonymous evolution to deviate considerably from those expected under neutrality. To investigate the quantitative relationship between parameters of mutation, selection, and demography, and patterns of synonymous site divergence, we have developed a novel combination of population genetic models and likelihood methods of phylogenetic sequence analysis. Comparing 50 orthologous gene pairs from *Drosophila melanogaster* and *D. virilis* and 27 from *D. melanogaster* and *D. simulans*, we show considerable variation between amino acids and genes in the strength of selection acting on codon usage and find evidence for both long-term and short-term changes in the strength of selection between species. Remarkably, *D. melanogaster* shows no evidence of current selection on codon usage, while its sister species *D. simulans* experiences only half the selection pressure for codon usage of their common ancestor. We also find evidence for considerable base asymmetries in the rate of mutation, such that the average synonymous mutation rate is 20–30% higher than in noncoding regions. A Bayesian approach is adopted to investigate how accounting for selection on codon usage influences estimates of the parameters of mutation.

IN phylogenetic and population genetic analyses, mutations at synonymous codon positions are typically considered to behave according to the neutral model of molecular evolution (Kimura 1983). However, in many species, including numerous prokaryotes (Sharp and Li 1986), unicellular eukaryotes (Sharp *et al.* 1986), and multicellular eukaryotes such as Drosophila (Shields *et al.* 1988), *Caenorhabditis elegans* (Stenico *et al.* 1994), and *Arabidopsis thaliana* (Chiapello *et al.* 1998; Duret and Mouchiroud 1999), there is considerable evidence that codon usage is influenced by selection acting at the level of translation. In particular, the relationship between transfer RNA abundance, codon usage, and gene expression level (Grantham *et al.* 1981; Ikemura 1981) suggests that the rate or accuracy with which alternative codons are translated has a significant influence on organismal fitness (Bulmer 1991).

Selection acting on synonymous codon positions affects not only base composition but also patterns of polymorphism and divergence (Hartl *et al.* 1994; Akashi 1995; Eyre-Walker and Bulmer 1995; McVean and Charlesworth 1999). Indeed, much of the evidence in support of selection has been derived from patterns of polymorphism and divergence at synonymous sites. Of particular importance has been the observation in both bacteria and eukaryotes that genes with higher codon bias have lower rates of substitution (Sharp and Li 1987; Shields *et al.* 1988). This agrees with the notion that when selection on codon usage is strong, most amino acids are encoded for by preferred codons, so the majority of new mutations are deleterious and will be effectively removed by selection.

A second consequence of selection on codon usage is that changes in the strength or efficacy of selection acting on codon usage may lead to unusual patterns of divergence. For example, an excess of substitutions to unpreferred codons in the lineage leading to *Drosophila melanogaster* from its most recent common ancestor (MRCA) with *D. simulans* has been interpreted in terms of a recent reduction in the effective population size (N_e) of *D. melanogaster* (Akashi 1996). In another case, an excess of substitutions to unpreferred codons in the *yellow* gene has been associated with a change in the recombinational environment (Takano-Shimizu 1999), in support of the idea that the efficacy of selection is reduced in regions of low recombination (Kliman and Hey 1993).

Such deviations from the neutral model of molecular evolution will clearly introduce biases into methods of phylogenetic analysis that assume that synonymous mutations are of no consequence to organismal fitness. The quantitative nature of the relationship between parameters influencing codon usage and patterns of divergence can be investigated through the use of explicit population genetics models (Bulmer 1991; Hartl *et al.* 1994; Akashi 1995; Akashi and Schaeffer 1997; McVean and Charlesworth 1999). Models based on the diffusion theory of Wright (1931) and Kimura

Corresponding author: Gilean McVean, Department of Statistics, 1 S. Parks Rd., Oxford OX1 3TG, United Kingdom.
E-mail: g.mcvean@ed.ac.uk

(1983) can be manipulated to estimate parameters of selection and mutation from codon frequencies (Bulmer 1991; McVean and Vieira 1999) and patterns of polymorphism and divergence (Hartl *et al.* 1994; Akashi 1995; Akashi and Schaeffer 1997). While these models make a number of simplifying assumptions, such as constant population size, the infinite-sites model (Kimura 1971), and evolutionary independence of the fates of simultaneously segregating mutations (McVean and Charlesworth 2000), under biologically realistic conditions they can provide reasonably accurate estimates of the underlying parameters (Akashi 1999; McVean and Charlesworth 2000).

In this article, we present a new approach to analyzing patterns of molecular evolution and specifically patterns of synonymous site divergence. The idea is to use conventional likelihood methods of phylogenetic sequence analysis, but to parameterize the underlying models of sequence evolution in terms of the population genetics of mutation, selection, and drift. The benefit of using models constructed simply from biologically meaningful parameters is that we can use patterns of divergence to provide estimates of the parameters of interest. The benefit of a likelihood approach is that we can both test whether increasingly complex models provide significantly better fits to the data and get an idea of the range of parameter values compatible with the data.

Population genetic models incorporating selection have previously been applied to patterns of synonymous-site evolution (Hartl *et al.* 1994; Akashi 1995, 1996), but these have relied on outgroup methods to assign the direction of mutations at synonymous sites through parsimony. However, parsimony can lead to inaccurate results arising from multiple hits, particularly when there is considerable heterogeneity between sites in the rate of substitution. In addition, it means that only very closely related species can be compared. In contrast, a phylogenetic method that corrects for multiple hits can be used to compare species of any divergence time.

The second major advantage of the method is that we can use information from both patterns of codon usage and patterns of synonymous divergence. Population genetic models predict not only substitution rates but also equilibrium codon frequencies. Data on codon frequencies have been previously used to analyze patterns of codon usage (Bulmer 1991; McVean and Vieira 1999), but no attempt has been made to combine these with divergence data.

In the following section, we describe a simple model for sequence evolution that includes parameters of mutation, selection, and demography. We then describe how this can be applied to data on synonymous site divergence and use it to analyze patterns of evolution in 50 homologous genes from *D. melanogaster* and *D. virilis* and 27 homologous genes from *D. melanogaster* and *D. simulans*. Simultaneous analysis of multiple genes lends much power to the approach and provides a way of testing for heterogeneity between genes in patterns of divergence. To this end, we develop a series of goodness-of-fit tests to consider the extent to which the model provides an adequate description of the data.

A POPULATION GENETIC MODEL FOR CODON USAGE EVOLUTION

The mutation-selection-drift model for the evolution of codon usage provides a simple population genetic description of the forces influencing synonymous site evolution (Bulmer 1991; McVean and Charlesworth 1999). Consider an amino acid with four codons, such as alanine (Figure 1), and assume that the rate of nonsynonymous substitution is negligible compared to the rate of synonymous substitution. Synonymous mutations appear at each codon type with a frequency determined by the current codon and the type of change. If codons have different relative fitnesses (a function of the speed and accuracy with which they are translated) and selection is genic (*i.e.*, heterozygotes are of intermediate fitness), the probability of fixation of a single new mutation in a diploid population with an effective population size of N_e and a census population size of N is given by the formula of Kimura (1962),

$$u(s) = \frac{2s(N_e/N)}{1 - e^{-4N_e s}}, \qquad (1)$$

where s is the selective advantage of the mutant allele over the ancestral allele and is negative if the mutation is deleterious.

If the rate of mutation per site per generation from codon type i to j is μ_{ij}, the number of new mutations entering the population each generation at a site fixed for codon i is a Poisson distribution with mean $2N\mu_{ij}$. For $2N\mu_{ij} \ll 1$ and $u(s) \ll 1$, the fate of a new mutation

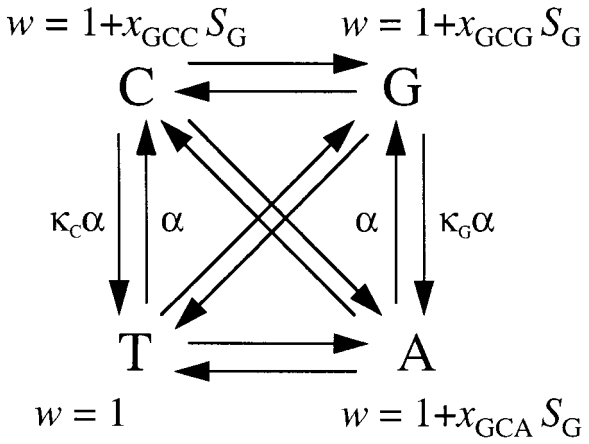

Figure 1.—Model of codon usage evolution for alanine: the relative fitness of codon i is a function of the hierarchy of selection x_i and the strength of selection on the gene S_G. The parameters of mutation estimated in the model are also shown.

is essentially independent of others occurring at the same locus either in the same or subsequent generations. If selection acts independently at each site in the genome, the rate of substitution from codon i to codon j is

$$\mu_{ij}\lambda_{ij}, \quad \text{where } \lambda_{ij} = \frac{S_{ij}}{1 - e^{-S_{ij}}} \tag{2}$$

and S_{ij} is $4N_e s_{ij}$ ($4N_e$ times the difference in fitness between alleles, scaled to an average fitness of one). When $S = 0$, this reduces to the mutation rate.

To use this population genetic model as a description of sequence evolution we must make the simplifying assumption that substitution events are instantaneous on an evolutionary timescale. That is, all differences observed between genes sampled from different species are the result of fixations governed by Equation 2 and are not the result of mutations segregating in either population. This assumption is implicit to almost all phylogenetic models of sequence evolution (see, *e.g.*, Zharkikh 1994) and is reasonable if between-species divergence is considerably greater than within-species polymorphism.

For an amino acid with n synonymous codons, and using a continuous time approximation, the substitution probabilities determine an $n \times n$ rate matrix $\mathbf{Q}$ that characterizes the rates of change in state for each codon. For example, an amino acid encoded for by two C- and T-ending codons, such as asparagine, will have the matrix

$$\mathbf{Q} = \begin{array}{c} \text{from/to} \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cc} C & \quad T \end{array} \\ \begin{bmatrix} -\lambda_{CT}\mu_{CT} & \lambda_{CT}\mu_{CT} \\ \lambda_{TC}\mu_{TC} & -\lambda_{TC}\mu_{TC} \end{bmatrix} \end{array}, \tag{3}$$

which is time independent if the parameters of mutation, selection, and demography (N_e) are invariant. The population size of a species is likely to vary over time, but if the timescale of fluctuations is short relative to the substitution process, N_e is well approximated by the harmonic mean population size (Crow and Kimura 1970, p. 110). In practice, we assign a separate N_e to each branch of the phylogenetic tree.

To describe patterns of sequence evolution, consider the transition probability matrix $\mathbf{P}_{(t)}$. For an amino acid encoded for by C- and T-ending codons this will have the form

$$\mathbf{P}_{(t)} = \begin{bmatrix} p_{CC} & p_{CT} \\ p_{TC} & p_{TT} \end{bmatrix}, \tag{4}$$

which describes the probabilities p_{ij} that a codon is of type j given that it was of type i, t generations ago. For a time-independent transition matrix with a continuous timescale approximation, this is given by

$$\mathbf{P}_{(t)} = e^{\mathbf{Q}t} \tag{5}$$

(see, *e.g.*, Jukes and Cantor 1969; Kimura 1980). We also use the rescaled time of species divergence, $\tau = \mu t$, where μ is the transversion mutation rate, as μ and t cannot be estimated separately.

For a given amino acid, the expected divergence matrix for homologous codons sampled at time τ after speciation, $\mathbf{D}_{(\tau)}$, where d_{ij} is the probability of observing a site with nucleotide i in species 1 and j in species 2, is given by

$$\mathbf{D}_{(\tau)} = [\mathbf{P}^1_{(\tau)}]^{\mathrm{T}}\mathbf{F}\mathbf{P}^2_{(\tau)}, \tag{6}$$

where $\mathbf{P}^1_{(\tau)}$ and $\mathbf{P}^2_{(\tau)}$ are the transition probability matrices for each species and $\mathbf{F}$ is a diagonal matrix of the frequencies of each codon in their common ancestor, given by the behavior of (5) as $\tau \rightarrow \infty$.

In the full model, N_e is estimated for each daughter species relative to the MRCA, such that the ratio $N_e(\text{species})/N_e(\text{MRCA})$ is given by the parameter f_{species}. Changes in N_e generate nonstationary patterns of substitution and asymmetrical divergence matrices. This approach also differs from current methods for estimating sequence divergence (*e.g.*, Yang and Nielsen 2000) in allowing nonreversible transition matrices, a frequent consequence of the influence of natural selection. The appendix presents a series of simulation results that show that under realistic conditions, the model provides reasonably accurate estimates of the parameters influencing synonymous site evolution.

APPLICATION TO EMPIRICAL DATA

The model outlined describes the behavior of codon usage for a given set of selection coefficients. But differences in codon usage between amino acids and genes should be allowed for when estimating parameters from data. As there is insufficient power to estimate the parameters of selection for each occurrence of each amino acid in each gene, our solution is to consider an explicit model for how the selection coefficient acting on codon usage varies between amino acids and genes (McVean and Vieira 1999). Specifically, we assume that while the overall strength of selection acting on codon usage varies between genes (for example, as a function of expression level), the relative strength of selection differentiating between codons does not. The hierarchy of selection on codon usage (McVean and Vieira 1999) is determined by assigning a relative fitness of 1 to one codon of each amino acid and a fitness of $1 + x_i S_G$ to other codons (Figure 1), where x_i is constant between genes and S_G is a parameter of the gene. For one amino acid (asparagine), the relative difference in fitness between C- and T-ending codons is fixed at 1, leaving a total of 39 parameters to be estimated. We assume that the relative strength of selection acting on orthologous genes has not changed between species, so that differences between species are due to genome-wide changes.

The choice of a linear scaling is simple but arbitrary and requires further investigation.

The power gained by these assumptions is that we can combine information from different genes, without assuming complete uniformity of the pressures affecting codon usage across the genome. This flexibility is considerably more realistic than previous approaches, but at the same time provides considerable power both to estimate the parameters determining codon usage and test the models against empirical data. In addition we can compare the fully parameterized model against simpler models through means of likelihood-ratio tests. We also use a simplified mutation model in which we estimate 4 of the 12 different mutation rates: transversions, rate μ; A $\rightarrow$ G/T $\rightarrow$ C transitions, rate $\alpha\mu$; G $\rightarrow$ A transitions, rate $\kappa_G\alpha\mu$; and C $\rightarrow$ T transitions, rate $\kappa_C\alpha\mu$ (see Figure 1). Models with more parameters do not provide a better fit to the data. Variation in the mutation rates between genes or sites is not considered.

We have applied the models to two sets of data concerning codon usage evolution in Drosophila. The first consists of 50 orthologous genes from *D. melanogaster* and *D. virilis* (McVean and Vieira 1999), thought to be separated by ~40 MY (Kwiatowski *et al.* 1994; Russo *et al.* 1995). The second is a set of 27 orthologous genes from *D. melanogaster* and *D. simulans*, separated by ~3 MY (Powell 1997). Patterns of codon usage and synonymous site divergence have been previously studied for both these comparisons (Moriyama and Gojobori 1992; Akashi 1996; McVean and Charlesworth 1999); however, our interests lie in understanding the extent to which such patterns can tell us about the underlying parameters of mutation, selection, and demography.

Sequences: Only complete gene sequences were used in the comparisons. Gene sequences for the *D. melanogaster-D. virilis* comparison were as previously used (McVean and Vieira 1999). For the *D. melanogaster-D. simulans* comparison we used 27 genes. For 12 of these, an outgroup sequence was available from within the *melanogaster* subgroup. Accession numbers are in the order *D. simulans/D. melanogaster/D. yakuba* (*D. orena* for *spalt*; *D. erecta* for *ref2p*); a semicolon indicates that multiple entries were used to construct the complete coding sequence: *achaete* (X62400/AH000975/3618324), *Acp26Aa* (X70899; X69686/4456056), *Acp26Ab* (X70899; X69686/4456056), *alcohol dehydrogenase* (M19263/Z00030/9239), *α-amylase distal* (D17733/L22720/413896), *α-amylase proximal*, (D17734/L22735/413898), *amylase-like protein* (U96159/U69607/AF039561), *cecropin A1* (AB010790/3192094), *cecropin B* (AB010790/AF018999), *decapentaplegic* (U63854/M30116), *esterase-6* (L10670/M15961), *fat body protein 2* (AF045786/AF045796), *glucose dehydrogenase* (U63325/M29298), *lethal-of-scute* (AB005802/298089/3618332), *myosin light chain* (L08051/K01567), *metallothionein N* (M69016/M69015), *cytochrome P450* (AF017005/AF017002), *glucosephosphate isomerase* (L27552/U20567/L27684), *ras 1* (AF186649/K01960), *ras3* (AF186655/8407), *ref2p* (U23930/8420/7407), *spalt* (M21227/8536/M21579), scute (AB005801/AH000975/3618330), *Cu-Zn superoxide dismutase* (X15685/8644/AF127159), *triosephosphate isomerase* (U60861/10945/U60870), *vermilion* (U27204/M34147) and *white* (U64875/10873). Protein sequence alignments were constructed for each gene pair, and only conserved amino acid positions were used in the analysis. Conserved amino acid positions are under stronger selection for codon usage in *D. melanogaster* (Akashi 1994), but while this is reflected in our results, it should not introduce any bias into the method. Average protein identity at alignable positions between *D. melanogaster* and *D. virilis* is 76% (range 37–97%). For *D. melanogaster* and *D. simulans*, average protein identity is 97% (range 73–100%). Average synonymous divergence (Nei and Gojobori 1986) between *D. melanogaster* and the outgroup sequences used to infer the direction of mutations through parsimony is 16% for *D. yakuba* (10 genes) and 20 and 34% for the single *D. erecta* and *D. orena* sequences.

Likelihood estimation: In the implemented version of the model, there are 45 parameters common to all genes: the time of species divergence, the ratio of current to ancestral population size for each species, 3 mutational parameters, and 39 parameters for the hierarchy of selection among codons. We treat serine as two separate amino acids because the two relevant sets of codons cannot be bridged in a single point mutation. For each gene the relative strength of selection on codon usage is also estimated.

The likelihood of the data for a given set of parameter values is calculated in two parts. For a given set of parameter values we generate a surface of expected codon divergence matrices (Equation 6) for each amino acid for a fixed number of classes of the strength of selection in the MRCA (21 classes evenly arranged over the interval $4N_e s = 0$–4, where the value of $4N_e s$ is that distinguishing between the codons AAC and AAT for asparagine). Codon divergence tables for each gene are compared to the expected frequencies and the likelihood is calculated as

$$L = \sum_i C_i \ln(f_i), \tag{7}$$

where C_i and f_{ij} are the observed count and expected frequency of the ith pair of codons. For each gene we find the class with the highest likelihood and the marginal maximum likelihoods are summed across genes. While we cannot guarantee to find the true maximum-likelihood value for the strength of selection, the error should be small for a sufficiently fine grid. We have also tried fitting a discrete gamma distribution to the selection coefficients; results obtained by this method are essentially identical to those presented here.

To explore the likelihood surface for the common parameters, we have adopted a Monte Carlo Markov chain (MCMC) approach, using the Metropolis-Hastings rejection algorithm. A parameter is chosen at random and incremented by a pseudorandom number drawn from a uniform distribution over the range $(-\delta, \delta)$, where δ has been previously fixed. The overall likelihood is then compared to the previous likelihood. If it is greater (*i.e.*, more likely), the change is accepted. If it is less likely, then the change is accepted with probability

$$\Pr\{\text{Accept}\} = \exp[L_{\text{new}} - L_{\text{old}}]$$

through comparison to a pseudorandom number drawn from a uniform distribution (0, 1). For reasonably smooth likelihood surfaces, this approach will explore parameter space in proportion to the likelihood, so samples from the MCMC can be used to obtain the posterior distribution of each parameter (with the assumption of a uniform prior for each parameter).

Although the Metropolis-Hastings algorithm will explore parameter values in proportion to their likelihood, two practical issues must be addressed. First, this behavior only applies once the MCMC reaches equilibrium, which takes an unknown period of time to achieve. Second, we also wish to find the maximum-likelihood (ML) parameter values (to compare models), and with such high dimensionality the MCMC takes a long time to find the ML values. For these reasons, we have split the problem in two. First, we use a form of simulated tempering to find the ML values. The MCMC is run as before, but for proposed changes with lower likelihood, the difference in log-likelihood is multiplied by a factor c (range 1–50). From the starting parameter values, the chain rapidly increases in likelihood. After a number of iterations (typically 5000 proposed changes), we then reduce c to <1 (range 0–0.5) for a number of iterations (typically 1000) and then revert to the original value. This is repeated a number of times and with different starting points, until the same ML values are found repeatedly. The parameter values arrived at by this method are taken to be the ML estimates. The MCMC is then run for 5×10^6 proposed changes, with samples taken every 500 proposed changes, after a burn-in of 5×10^5 proposed changes, to estimate the posterior distribution of each parameter. The procedure was repeated several times to check for convergence.

RESULTS

We first consider whether the addition of extra parameters of selection and demography significantly improves the fit between model and data. Table 1 shows the increase in log-likelihood for a series of improvements to a basic model in which all synonymous mutations are neutral. The major codon usage (MCU) selection model assumes that the difference in fitness between preferred and unpreferred codons is the same for all amino acids in each gene, but that there is variation between genes in the strength of selection acting on codon usage. Preferred codons are those defined by Akashi (1995). The codon model includes the hierarchy of selection among codons discussed earlier. The lineage N_e model allows each daughter species to have a separate N_e, which may be different from that of their ancestor. Each model improvement in both comparisons provides a significant increase in likelihood, as assessed by the approximation that twice the difference in log-likelihood is approximately χ^2 distributed with the degrees of freedom equal to the difference in the number of parameters (Sokal and Rohlf 1995, p. 689).

Relative mutation rates: We use a simplified mutation model that considers four different mutation rates (see Figure 1). The reason for separating different types of transition is that there is considerable evidence for an AT-biased mutation process in Drosophila (Kliman and Hey 1994), and, as transition mutations are more common than transversions (Moriyama and Powell 1996), the bias is perhaps most likely to result from differences in transition rates. We find strong evidence for a considerable AT mutation bias at transitions for both comparisons, such that the rates of C → T and G → A mutations are 1.5–3 times higher than T → C and A → G mutations (Table 2). In both comparisons the maximum *a posteriori* (MAP) parameter estimates are considerably lower than the ML estimates, but included within the 2-unit support intervals.

We also find evidence for an elevated rate of other transitions: A → G and T → C transitions are estimated to be one to two times as common as transversions in the two comparisons. Both sets of values predict that in regions under no selection on base composition, the GC content should be 39–40%. This compares with empirical estimates from intron sequences in which the average GC content is ~37% (Kliman and Hey 1994).

TABLE 1

Models and likelihoods

Model	No selection	MCU selection	Codon model	Lineage N_e
mel-vir	$L^a = -39{,}871$ (4)	$\Delta L = 959$ (54)	$\Delta L = 1{,}699$ (93)	$\Delta L = 45$ (95)
mel-sim	$L = -11{,}229$ (4)	$\Delta L = 667$ (31)	$\Delta L = 735$ (70)	$\Delta L = 46$ (72)

[a] Log-likelihood of model (number of parameters). ΔL is the increase in log-likelihood from the previous model.

TABLE 2

Parameter estimates

	mel-vir		*mel-sim*	
Parameter	ML	MAP (95% CI)	ML	MAP (95% CI)
τ^a	0.83	0.87 (0.81–0.92)	0.047	0.051 (0.046–0.058)
κ_C	3.5	2.2 (2.0–2.5)	2.3	1.7 (1.4–2.1)
κ_G	2.7	1.9 (1.6–2.4)	2.5	1.8 (1.4–2.4)
α	0.88	1.1 (0.95–1.3)	1.5	1.8 (1.4–2.3)
f_{mel}	0.45	0.47 (0.35–0.59)	0.093	0.10 (−0.088–0.21)
f_{vir}	0.40	0.41 (0.29–0.51)	—	—
f_{sim}	—	—	0.53	0.53 (0.33–0.68)

ML, maximum likelihood; MAP, maximum *a posteriori*; CI, credibility interval.
[a] The time since divergence expressed as the expected number of mutations per site in noncoding regions.

In addition, these values predict that in noncoding regions 44–54% of all new mutations should be transitions. This compares to values of 54% in *D. melanogaster* and 47% in *D. simulans* (Moriyama and Powell 1996), estimated from patterns of polymorphism.

The hierarchy of selection on codon usage: To describe the influence of selection on codon usage, we estimate the fitness of codons relative to the others within the same synonymous group and the strength of selection relative to other amino acids: what we have previously called the hierarchy of selection on codon usage (McVean and Vieira 1999). Figure 2 compares the hierarchy expressed as the maximum difference in relative fitness between codons for each amino acid, for the two species-pair comparisons. Estimates from the two comparisons are highly correlated, as expected, given that one species is common to both. There are two notable features. First, there is considerable variation between amino acids in the strength of selection acting on codon usage, with aspartic acid experiencing the least and leucine the greatest. Second, an important determinant of the strength of selection is the number of synonymous codons (*mel-vir*, Pearson correlation coefficient $\rho = 0.73$, $P = 0.004$ by randomization; *mel-sim*, $\rho = 0.57$, $P = 0.01$); see also Kliman and Hey (1994).

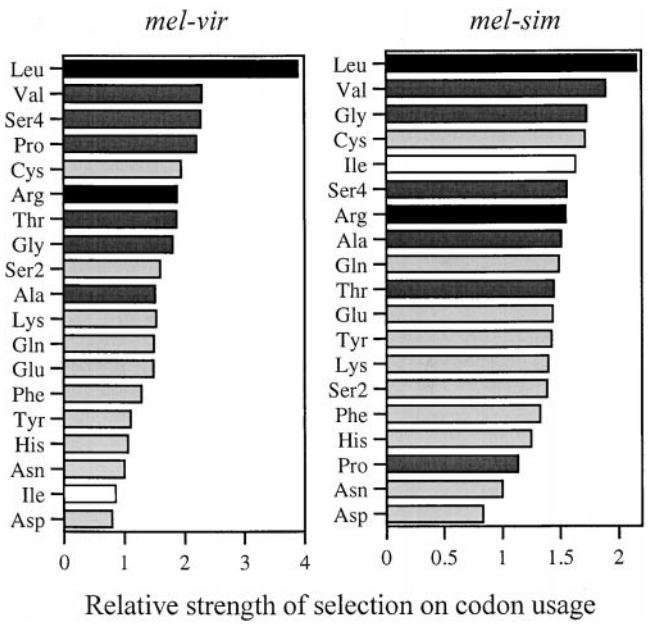

Figure 2.—The ML estimate of the hierarchy of selection acting on codon usage from each comparison. Values represent the maximum difference in fitness between codons for each amino acid, relative to that of asparagine. Amino acids with two (gray), three (white), four (dark gray), and six (black) codons are shown.

There are two possible explanations. With more codons and more tRNAs, there is greater opportunity for unfavorable tRNA-codon pairings and consequently greater selection for codons that correspond to the more abundant tRNAs. Alternatively, given that there is a correlation between the relative frequency of amino acids and the number of codons ($\rho = 0.53$, $P = 0.018$; $\rho = 0.58$, $P = 0.01$ for the two data sets), a correlation between codon number and strength of selection is expected if the frequency of an amino acid determines the strength of selection on codon usage. To distinguish these hypotheses, we can consider the relationship between amino acid frequency and strength of selection within each group of amino acids. We find that observed patterns vary between comparisons, but, if anything, there is a negative correlation between amino acid abundance and strength of selection on codon usage within groups: amino acids with two codons, $\rho = -0.09$, $P = 0.80$; $\rho = -0.57$, $P = 0.09$; four codons, $\rho = -0.85$, $P = 0.026$; $\rho = 0.68$, $P = 0.15$. Abundance *per se* is therefore not an important determinant of the strength of selection on codon usage. We suggest that amino acids with more codons have stronger selection on codon usage because they have a greater opportunity for unfavorable codon-tRNA pairing.

Demographic differences between species: For each species comparison we consider a simple demographic model in which an ancestral population at mutation-selection-drift equilibrium gives rise to two daughter species of variable population size. We estimate the

TABLE 3

Substitutions to preferred and unpreferred codons

	Preferred[a]	Unpreferred
D. melanogaster	17	95
D. simulans	24	47

[a] Substitutions are estimated by parsimony from outgroup sequences (*D. yakuba*, *D. orena*, and *D. erecta*).

strength of selection acting on codon usage in the ancestral population, the relative strength in each daughter species (f_{sp}), and the time since the speciation event in terms of the expected number of mutations in neutral noncoding regions (Table 2).

As previously (McVean and Vieira 1999), we find evidence for a lower strength of selection on codon usage in the lineage leading to *D. virilis* than that leading to *D. melanogaster*. However, in contrast to our previous result of a relative strength of only ~50%, based on amino acids with two codons, the relative strength in *D. virilis* is estimated to be ~90% that of the *D. melanogaster* lineage, and both experience less than half the strength of selection on codon usage than their MRCA.

In the *mel-sim* comparison, we find that both species have a lower strength of selection on codon usage than their MRCA. But whereas *D. simulans* has about one-half the selection of the ancestral species, selection in *D. melanogaster* has been reduced to one-tenth of its ancestral level and is not significantly different from zero. A decrease in the strength of selection ($N_e s$) in *D. melanogaster* since divergence from its MRCA with *D. simulans* has been previously suggested from patterns of molecular evolution (Akashi 1996). This may have been caused by a protracted reduction in N_e, resulting from a bottleneck, or strong population subdivision with nonconservative migration. The decrease in N_e for *D. simulans* has not been previously detected. This conclusion is also reached if we use the same outgroup approach as Akashi (1995) but with a larger sample size (12 genes). Table 3 shows that there is a significant apparent excess of mutations to unpreferred codons ($\chi^2_{\nu=1} = 4.07$, $P < 0.05$) in the *D. simulans* lineage (preferred codons classified according to Akashi 1995).

Testing the adequacy of the model: The methods presented here provide a powerful approach to estimating parameters of mutation, selection, and demography from patterns of synonymous site evolution. However, that the parameter values we estimate are only as good as the model is an accurate description of the data. There are two ways in which our model may be inaccurate. First, the population genetic model makes assumptions that are not valid (particularly independence between sites and the neglect of polymorphism). These are discussed in the appendix. Second, the constraints we impose to apply the model to sequence data may be inaccurate. For example, we do not consider variation in the selection coefficient between occurrences of an amino acid within a gene or variation in the mutation rate between genes. In short, the number of parameters in the model is bound to be fewer than in reality.

To test the adequacy of the model as providing an explanation of the data, we consider a series of goodness-of-fit tests. The simplest test is to compare the observed number of each codon pair with that expected from the ML parameter values. We calculate a *G*-test statistic as the sum over all possible homologous codon pairs *i*, amino acids *j*, and genes *k*,

$$G_{tot} = 2\sum_{ijk} C_{ijk} \ln(C_{ijk}/E[C_{ijk}]),$$

where C_{ijk} is the observed number of such pairs and $E[C_{ijk}]$ is the expected number (Sokal and Rohlf 1995, p. 690). This statistic is compared to a null distribution obtained by simulating 10,000 data sets using the ML parameter values. For both comparisons, G_{tot} is greater than all simulated values; hence we can reject the hypothesis that the model provides an adequate description of the data. For the *mel-vir* comparison 27 out of the 50 genes have G_{tot} values >95% of simulated sequences, and for the *mel-sim* comparison the proportion is 9 out of 27. In short, for between one-third and one-half of the genes the model does not provide an adequate fit.

What is the cause of the discrepancy? There are two important ways in which the model may fail. First, the level of divergence may vary more between genes and amino acids than the model predicts. Second, deviation between observed and expected codon frequencies may result from the constraints imposed by the fixed hierarchy of selection. To investigate discrepancies between the model and data in the level of divergence, we consider a *G*-test statistic obtained by comparing the observed (D_{jk}) and expected numbers of homologous codons at which there are differences, across amino acids *j* and genes *k*:

$$G_d = 2\sum_{jk} D_{jk} \ln(D_{jk}/E[D_{jk}]).$$

The significance level of each test is calculated by comparing the observed value to those from 20,000 simulated data sets.

In both comparisons, a number of genes show strong deviation from the expected divergence (nine and six, respectively), although this is not consistently in one direction. We also consider G_d for each amino acid by summing the observed and expected levels of divergence for each amino acid in each gene. For 7 (*mel-vir*) and 3 (*mel-sim*) of the 19 amino acids, the model predicts levels of divergence significantly different from those observed.

To investigate the extent to which the model provides an adequate fit to observed codon frequencies in each species, we consider a *G*-test statistic obtained by sum-

ming the discrepancy between observed and expected codon counts, over codons i, amino acids j, and genes k:

$$G_{\text{cod}} = 2\sum_{ijk} C_{ijk} \ln(C_{ijk}/E[C_{ijk}]).$$

Note that this test does not consider divergence, and codon frequencies in each species are considered separately. We also partition this value by genes and amino acids. Significance is assessed by comparison to 10,000 data sets simulated from the ML parameter estimates.

We find significant deviations from the model for both genes and amino acids. In the *mel-vir* comparison, 18 (*mel*) and 25 (*vir*) genes have G_{cod} values >95% of simulated values, while for the *mel-sim* comparison, the numbers are 3 (*mel*) and 5 (*sim*) genes. Partitioning by amino acids we find that most amino acids show significant deviation in the *mel-vir* comparison, although only 3 (*mel*) and 5 (*sim*) do in the *mel-sim* comparison. In short, the simple model is adequate for the majority of genes and amino acids in terms of describing levels of sequence divergence, but there is considerable discrepancy between the model and data in terms of the fitted codon frequencies for the more divergent species pair. A possible cause of the discrepancy is changes in the relative strength of selection acting on codon usage between amino acids over the timescale separating *D. melanogaster* and *D. virilis* (McVean and Vieira 1999).

DISCUSSION

The method presented here is the first attempt to combine conventional phylogenetic methods of sequence analysis with populations genetic models of the underlying substitution process, incorporating mutation, selection, and drift. By applying these ideas to patterns of synonymous site divergence in Drosophila, we have estimated the strength of selection acting on codon usage and detected demographic events in the history of the species that have had a profound effect on the efficacy of selection. We suggest that this approach provides a powerful method for understanding patterns of sequence evolution because the underlying models are constructed exclusively from biologically meaningful parameters.

The effect of base composition on the mutation rate: Our analysis provides strong evidence for considerable mutational bias in Drosophila, such that the rates of C → T and G → A mutation are about two times higher than the reverse. This has important consequences for interpreting patterns of molecular evolution and variation, because it implies a link between the base composition of a region and its mutation rate. This point has been made before (McVean and Charlesworth 1999), but we have not previously been able to quantify the difference. From the codon frequencies observed in our data sets and the ML estimates of the mutation parameters, we predict that $\bar{\mu}_S$, the average synonymous mutation rate (in terms of synonymous mutations per synonymous site as estimated by the method of Nei and Gojobori 1986), should be 20–30% higher on average than the noncoding mutation rate, μ_{NC}, for *D. melanogaster* (Table 4). For highly biased genes, the synonymous mutation rate may be 40% higher than the noncoding mutation rate. In part this is because the method of Nei and Gojobori (1986) does not account for the higher frequency of transition mutations relative to transversions. However, if just fourfold degenerate positions are considered we still expect a 10% higher mutation rate, $\bar{\mu}_4$, at synonymous positions. Almost identical values are obtained if we use a set of 2070 genes (EDGP 1999) to obtain average codon frequencies (Table 4).

A consequence of a higher synonymous mutation rate is that we expect synonymous site polymorphism to be greater than that in noncoding regions, as long as the strength of selection acting on codon usage is relatively weak (McVean and Charlesworth 1999). We have shown that the strength of selection acting on codon usage in *D. melanogaster* is not significantly greater than zero, hence in this species we expect synonymous site diversity to be higher than that in noncoding regions. In addition, we predict a correlation between the level of polymorphism and the GC content of genes. Both patterns have been observed in *D. melanogaster* (Moriyama and Powell 1996), and 30–90% higher levels of synonymous *vs.* noncoding polymorphism have also been described in *D. simulans* and *D. pseudoobscura* (Moriyama and Powell 1996). In *D. virilis* the opposite pattern is found, with average intronic variation being ~20% higher (Vieira and Charlesworth 1999).

An alternative explanation for higher variability at synonymous sites is that introns and other noncoding regions of the genome are under stronger constraint than synonymous sites. To test this, we can compare the level of divergence observed in introns with that predicted by our best-fitting model under the assumption of neutrality. For the *mel-sim* comparison, the ML values predict that the average proportion of differences between homologous sequences in neutral noncoding regions should be 8.9%. From an analysis of 30 homologous internal introns (within the open reading frame) from 21 genes in *D. melanogaster* and *D. simulans* we find an average difference of 8.7% (274 changes in 3154 alignable sites, excluding GT . . AG motifs). While this analysis is not conclusive evidence for a lack of constraint in introns, it is compatible with the majority of changes in intron sequences being neutral. Higher synonymous than noncoding diversity in *D. melanogaster* and *D. simulans* may therefore simply be due to a higher mutability of these sequences. For *D. virilis* it may be that selection on codon usage is sufficiently strong to reduce synonymous polymorphism below that in introns. However, Vieira and Charlesworth (1999) did not detect skews in the allele frequency spectrum of

TABLE 4

Estimated parameters of mutation in *D. melanogaster*

	mel-vir comparison		*mel-sim* comparison	
	mel data	EDGP 2070 genes	*mel* data	EDGP 2070 genes
$\overline{\mu}_S/\mu_{NC}$	1.24 (1.20–1.29)[a]	1.23 (1.19–1.26)	1.33 (1.23–1.40)	1.3 (1.22–1.38)
$\overline{\mu}_4/\mu_{NC}$	1.11 (1.09–1.14)	1.10 (1.07–1.11)	1.10 (1.06–1.14)	1.10 (1.04–1.12)
U_S	0.009 (0.005–0.018)	0.01 (0.006–0.021)	0.005 (0.003–0.012)	0.011 (0.006–0.021)
U_N	0.024 (0.014–0.054)	0.028 (0.016–0.057)	0.024 (0.015–0.052)	0.028 (0.016–0.058)

[a] Maximum *a posteriori* estimate (95% credibility interval).

synonymous mutations as would be indicative of selection.

Estimating parameters of mutation, selection, and drift from patterns of molecular evolution: Using sequence data to estimate parameters such as the mutation rate and the strength of selection influencing new mutations has many advantages, most importantly the ease of data collection. But the accuracy and efficacy of such methods is debatable. The factors affecting molecular evolution are clearly far more complex than any model that can be fitted, and naive model comparisons have the potential to lead to greater faith in parameter estimates than the data truly allow.

A partial solution is to consider the ability of a model to provide an adequate explanation of the data before using it to obtain parameter estimates. The inability to reject a model does not guarantee that the model is correct, but parameters estimated from models that can be rejected should be treated with caution. The problem with this approach is that the factors affecting molecular evolution are so diverse that as the complexity of the fitted model increases, so may the power to reject it. For example, the model we have considered here is by a long way the most complex yet analyzed, but we have also shown that the model is inadequate in several important respects. While we do not wish to claim that all our parameter estimates are insensitive to such inadequacy, it may well be that some, or even the majority, are. No general statement can be made about which parameters we expect to be strongly biased by model inaccuracy; this can only be addressed through simulation (see the APPENDIX).

The second limitation in using molecular evolutionary analysis is that often the models do not estimate the parameters of interest, but only some compound parameter such as $t\mu$ or $N_e s$. To estimate the per-generation mutation rate, or the selection coefficient differentiating between alternative codons, we therefore have to obtain estimates of other parameters. Often no particularly reliable estimate of such nuisance parameters is available.

One solution is to use conservative, or extreme, estimates to define upper and lower bounds. This approach not only throws away information but also leads to parameter estimates that cannot be combined. For example, to estimate the per-generation neutral mutation rate in Drosophila, we could use the point estimates of 45 MY for the separation of the *D. melanogaster* and *D. virilis* lineages, 3 MY for the separation of *D. melanogaster* and *D. simulans*, and an average of 10 generations per year in the wild (POWELL 1997). Combined with the ML estimates from the current analyses, we obtain point estimates of the per-generation neutral mutation rate of 1.8×10^{-9} for the *mel-vir* comparison and 1.1×10^{-9} for the *mel-sim* comparison. But we have no idea of which estimate to trust more, whether they are really different from one another, or how to combine the estimates.

An alternative solution is to adopt a Bayesian approach and estimate the posterior distribution of the parameters of interest. The posterior distributions of the model parameters are obtained from the Monte Carlo Markov chain analysis (though note the previous caveat about model adequacy). The choice of priors for the other parameters is subjective, and we should err on the side of caution. The split of the Drosophila and Sophophora subgenera has been put at no more recent than 30 MYA and perhaps as early as 60 MYA, while from biogeography the split between *D. melanogaster* and *D. simulans* is thought to be 2–4 MYA (POWELL 1997). In the lab, *D. melanogaster* and *D. simulans* can reach up to 25 generations a year, but in the wild, the average is probably closer to 10.

These priors used are represented in Figure 3, as are the resulting posterior distributions for each comparison and the combined posterior. The combined MAP estimate is 1.5×10^{-9} mutations per site per generation in noncoding DNA, and the 95% credibility interval is 10^{-9}–2.5×10^{-9}. The distributions obtained from the two comparisons are very similar, which perhaps implies that sequence data are not the limiting factor in the accuracy of our estimate. The single greatest element of uncertainty in the calculation is the number of generations per year. Remarkably, the MAP estimate is almost exactly that obtained by KEIGHTLEY and EYRE-WALKER (1999), using the average synonymous divergence between the *melanogaster* and *obscura* groups (LI 1997, p. 191) and assuming an average of 10 generations per year. Likewise, average synonymous divergence between

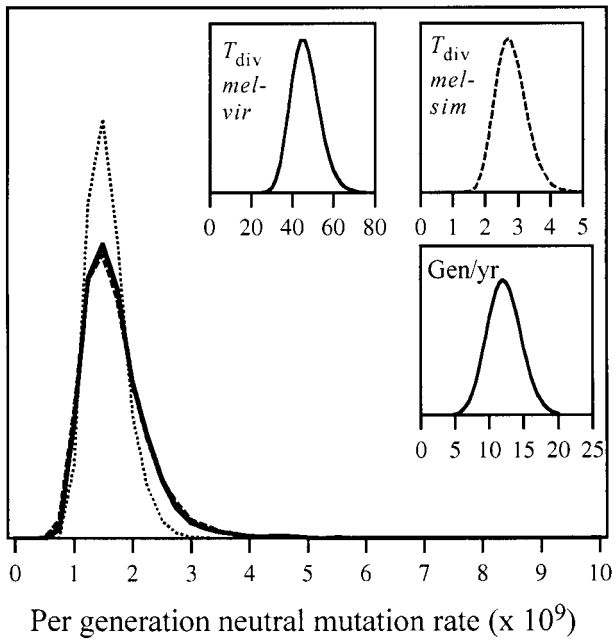

FIGURE 3.—The posterior distribution of the per-generation neutral mutation rate for the *mel-vir* (solid line), *mel-sim* (dashed line), and combined (dotted line) analyses. Also shown are the assumed priors for the time of divergence for each species comparison (in millions of years) and the number of generations per year.

D. melanogaster and *D. virilis*, calculated by the method of NEI and GOJOBORI (1986), predicts a rate of 1.4×10^{-9} mutations per site per generation (assuming 45 MY divergence) and divergence between *D. melanogaster* and *D. simulans* predicts a rate of 1.7×10^{-9} (assuming 3 MY since divergence).

Why should accounting for selection on codon usage have such a small effect on estimates of the per-generation mutation rate? Figure 4 shows, for each comparison, the relationship between the relative strength of selection acting on codon usage, the observed and expected proportions of differences at synonymous sites, and the differences expected at completely neutral sites. Two features are of note. First, the observed average proportions of synonymous differences are very close to the expected proportions of differences at noncoding sites. Second, the expected relationship between the strength of selection acting on codon usage and rate of synonymous divergence is fairly weak, particularly for the *mel-sim* comparison. In short, weak selection, of the order required to explain codon usage patterns in Drosophila, does not greatly influence the overall rate of substitution (though it does influence the type of mutations that become fixed between species).

The genome-wide rate of mutation: For several aspects of evolutionary theory, genome-wide parameters of mutation are more relevant than rates at individual sites. Specifically, the per-genome rate of deleterious mutation is a critical parameter in determining the relevance of one class of theory for the maintenance of sexual reproduction (KONDRASHOV 1988). Molecular evolutionary analysis can provide an estimate of the degree of constraint acting on DNA by comparing the rate of substitution with that of a region known to be evolving in a neutral fashion (KONDRASHOV and CROW 1993). Synonymous mutations are thought to be neutral in mammals and have been used to provide an estimate of the mutation rate in an analysis of hominid genes (EYRE-WALKER and KEIGHTLEY 1999).

This approach is clearly not valid in organisms for which there is selection acting on codon usage (KEIGHTLEY and EYRE-WALKER 1999). Indeed, mutations at synonymous sites will make a significant contribution to the deleterious mutation rate. Furthermore, because mutation biases create a link between base composition and the mutation rate, there is no simple relationship between the rate of synonymous evolution and the rate of nonsynonymous mutation.

Such complications are directly addressed by the methods presented here. To calculate the genome-wide rate of synonymous mutation in the *D. melanogaster* genome, U_S, we use the previous priors for the times of

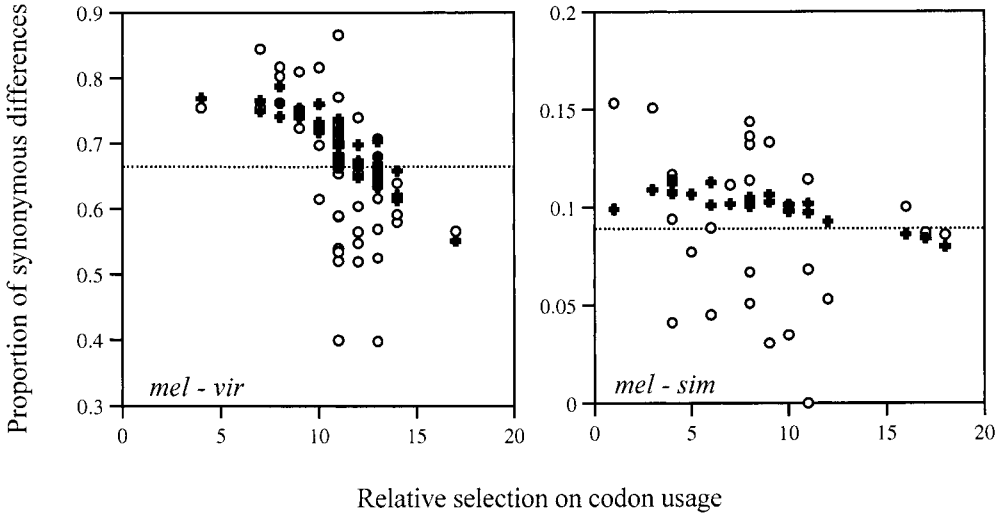

FIGURE 4.—The relationship between the relative strength of selection acting on codon usage and the observed (circles) and expected (crosses) proportions of synonymous differences in the two comparisons. The dotted line indicates the proportion of differences expected in noncoding DNA. Synonymous site differences are calculated by the method of NEI and GOJOBORI (1986); not corrected for multiple mutations.

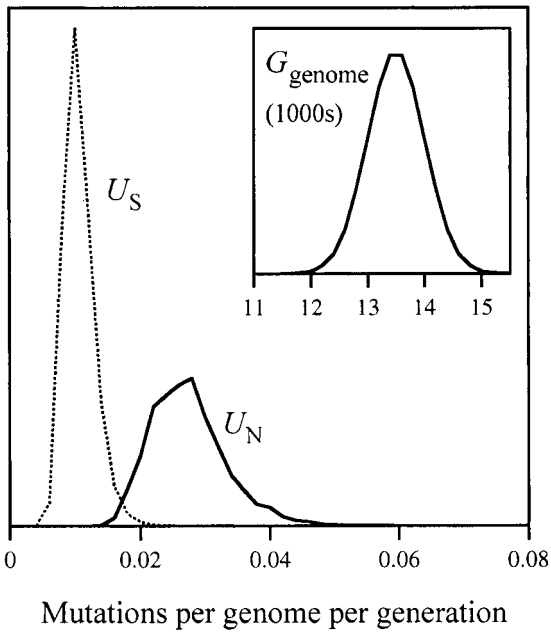

FIGURE 5.—The combined posterior distributions of the genome-wide synonymous and nonsynonymous mutation rates in *D. melanogaster*. Values are predicted from the codon frequencies in 2070 genes (EDGP 1999). Also shown is the assumed prior for the number of genes in the *D. melanogaster* genome.

divergence and the number of generations per year, and a prior for the number of genes in the *D. melanogaster* genome as shown in Figure 5 (G_{genome}: range 11–15 (ASHBURNER 1989, p. 104; ADAMS *et al.* 2000)). Combining the MCMC samples with the codon frequencies in the data sets, we obtain a MAP estimate of 7.5×10^{-3} synonymous mutations per haploid genome per generation (Table 4), of which about two-thirds are deleterious. The total contribution to the deleterious mutation rate from synonymous mutations is therefore $\sim 5 \times 10^{-3}$. However, from the *mel-sim* comparison we estimate that no synonymous mutations experience a selection intensity of $|N_e s| > 1$ in either *D. melanogaster* or *D. simulans*, and only 6% did in their MRCA. Using a data set of 2070 genes (EDGP 1999) to obtain codon frequencies gives a slightly larger combined MAP estimate of 0.011 synonymous mutations per haploid genome per generation, due to the underrepresentation of long genes in the samples (Figure 5).

The per-genome rate of nonsynonymous mutation, U_N, in *D. melanogaster* can be estimated by counting the number of mutations in genes that would lead to a change in the amino acid. We obtain a combined MAP estimate of 0.024 nonsynonymous mutations per haploid genome per generation from the data sets used here and an estimate of 0.028 from the EDGP data (Table 4 and Figure 5). Again, this value is almost identical to an estimate that did not take into account selection on codon usage (KEIGHTLEY and EYRE-WALKER 1999). Even if all nonsynonymous mutations were deleterious, the total deleterious mutation rate in *D. melanogaster* due to point mutations in coding sequences is very unlikely to be >0.075 per haploid genome per generation. Fitness-based estimates of the minimum deleterious mutation rate in *D. melanogaster* are in the range of 0.025–0.3 per haploid genome per generation (DRAKE *et al.* 1998). If the larger estimates are closer to the true value, point mutations in coding regions can constitute only a minor fraction of the total deleterious mutation rate.

The genetic load caused by synonymous mutations: While the selection coefficients associated with mutations at any one site are likely to be very small, the cumulative effects of a large number of unpreferred codons across the genomes may be considerable. KONDRASHOV (1995) has suggested that the accumulation of unpreferred codons presents a paradox in genetic load, such that the human genome may contain of the order of 100 lethal equivalents. The extent to which this is true depends on the strength of selection acting on codon usage, specifically, the product across genes

$$\prod_i (1 - s_i)^{C_i},$$

where s_i is the selective disadvantage of a codon relative to the best possible and C_i is the number of occurrences of that codon. From the results presented, we can estimate the genetic load caused by unpreferred codons in the *D. melanogaster* genome.

Using the approximation $(1 - s)^C \approx e^{-sC}$, the estimated genetic load in the genome as a whole is given by $L = 1 - \exp[(G_{genome}/G_{samp})\Sigma_i(w_i - w_{opt})C_i/4N_e]$, where the relative fitnesses of codons have been estimated from the model and G_{samp} and G_{genome} are the numbers of genes in the sample and the genome as a whole. Estimates of the recent N_e of *D. melanogaster* are $\sim 10^6$ (MORIYAMA and POWELL 1996). If we assume that the long-term N_e is the same as the recent N_e, this gives an estimated genetic load of 6%. This is probably an underestimate because the long-term N_e is likely to be smaller than the recent N_e. In short, unpreferred codons may impose a considerable genetic load in Drosophila, even though their contribution to standing variance in fitness is expected to be negligible.

We thank Brian Charlesworth, Adam Eyre-Walker, Peter Keightley, Arcadi Navarro, Rasmus Nielsen, Ziheng Yang, and an anonymous reviewer for discussion and comments on the manuscript. G.M. is funded by the Natural Environment Research Council, and J.V. is supported by the Fundação para a Ciencia e Tecnologia (PRAXIS XXI/BPD/14120/97).

LITERATURE CITED

ADAMS, M., S. CELNIKER, R. HOLT, C. EVANS, J. GOCAYNE *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

AKASHI, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

AKASHI, H., 1995 Inferring weak selection from patterns of polymor-

phism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 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.

Akashi, H., 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.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bulmer, M. G., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

Chiapello, H., F. Lisacek, M. Caboche and A. Henaut, 1998 Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene **209:** GC1–GC38.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Rowe, New York.

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

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila* and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

EDGP, 1999 European *Drosophila* Genome Project: nuclear_cds_set.embl.v2.2. ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/ARCHIVE.

Eyre-Walker, A., and M. Bulmer, 1995 Synonymous substitution rates in enterobacteria. Genetics **140:** 1407–1412.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature **397:** 344–347.

Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **8:** r49–r62.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:** 389–409.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by N. H. Munro. Academic Press, New York.

Keightley, P. D., and A. Eyre-Walker, 1999 Terumi Mukai and the riddle of deleterious mutation rates. Genetics **153:** 515–523.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713–719.

Kimura, M., 1971 Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. **2:** 174–208.

Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

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

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Kliman, R. M., and J. Hey, 1994 The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics **137:** 1049–1056.

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

Kondrashov, A., 1995 Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J. Theor. Biol. **175:** 583–594.

Kondrashov, A. S., and J. F. Crow, 1993 A molecular approach to estimating the human deleterious mutation-rate. Hum. Mutat. **2:** 229–234.

Kwiatowski, J., D. Skarecky, K. Bailey and F. J. Ayala, 1994 Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the *Cu, Zn Sod* gene. J. Mol. Evol. **38:** 443–454.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

McVean, G. A. T., and B. Charlesworth, 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. **74:** 145–158.

McVean, G. A. T., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

McVean, G. A. T., and J. Vieira, 1999 The evolution of codon preferences in Drosophila: a maximum likelihood approach to parameter estimation and hypothesis testing. J. Mol. Evol. **49:** 63–75.

Moriyama, E. N., and T. Gojobori, 1992 Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **130:** 855–864.

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

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Powell, J., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, Oxford.

Russo, C. A. M., N. Takezaki and M. Nei, 1995 Molecular phylogeny and divergence time of Drosophilid species. Mol. Biol. Evol. **12:** 391–404.

Sharp, P. M., and W.-H. Li, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Sharp, P. M., T. M. E. Tuohy and K. R. Mosurski, 1986 Codon usage in yeast-cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5143.

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

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman and Company, New York.

Stenico, M., A. T. Lloyd and P. M. Sharp, 1994 Codon usage in *Caenorhabditis elegans*—delineation of translational selection and mutational biases. Nucleic Acids Res. **22:** 2437–2446.

Takano-Shimizu, T., 1999 Local recombination and mutation effects on molecular evolution in Drosophila. Genetics **153:** 1285–1296.

Vieira, J., and B. Charlesworth, 1999 X chromosome DNA variation in *Drosophila virilis*. Proc. R. Soc. Lond. Ser. B **266:** 1905–1999.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Yang, Z., and R. Nielsen, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17:** 32–43.

Zharkikh, A., 1994 Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39:** 315–329.

Communicating editor: J. Hey

APPENDIX

The assumptions of the model: The underlying model of sequence evolution we have employed makes a number of critical assumptions. The two most important are first that the substitution process is instantaneous and, therefore, that segregating mutations do not contribute to differences between sequences from different species. The second is that evolution acts independently at all sites. Neither assumption is realistic, so it is important to understand the effects of relaxing these assumptions on parameter estimates.

To this end we have conducted a series of simulations

TABLE A1

Mean (SD) ML estimates from 50 simulated data sets

	Estimated time[a]				
Time	$4N_e s = 0$	$4N_e s = 1$	$4N_e s = 2$	$4N_e s = 4$	$4N_e s = 10$
0	0.019 (0.010)	0.019 (0.007)	0.022 (0.005)	0.023 (0.011)	0.088 (0.049)
0.05	0.068 (0.009)	0.071 (0.007)	0.072 (0.009)	0.084 (0.012)	0.166 (0.049)
0.1	0.119 (0.014)	0.124 (0.011)	0.122 (0.014)	0.137 (0.023)	0.208 (0.048)
0.2	0.221 (0.019)	0.226 (0.016)	0.229 (0.025)	0.245 (0.047)	0.211 (0.030)
0.5	0.533 (0.089)	0.542 (0.087)	0.588 (0.154)	0.554 (0.107)	—
Estimated $4N_e s$	0.02 (0.06)	0.87 (0.05)	1.53 (0.06)	2.54 (0.09)	4.85 (0.26)

[a] Excluding cases where $\hat{\tau} = \infty$.

in which we consider a more biologically plausible speciation model and a finite sample size. We consider a two-allele model with selection and symmetric, reversible mutation, and a population of 500 diploid individuals, each with a genome of 1000 linked sites. At speciation, we assume that the population is simply duplicated (each daughter species has the same population size). We sample a single sequence from each daughter population at different time points and find the ML estimates of the selection coefficient and time since divergence (we assume that the relative mutation rates are known). For each combination of selection coefficient and time we take 50 independent samples: the scaled parameter values used are $4N_e\mu = 0.04$, $4N_e r = 0.1$, and $4N_e s$ in the range 0–4.

Table A1 shows the results of the simulations. The main effect is that the estimated strength of selection tends to be lower than the true strength of selection, even when the recombination rate between adjacent sites is relatively high. For weakly selected sites the bias is small (10–20%), while for $4N_e s \geq 4$, the average downward bias can be at least 40%, although for lower values of $N_e\mu$, the effects are weaker (data not shown). Both ancestral polymorphism and that generated subsequent to species divergence tend to lead to overestimates of the time since divergence, and for sites under strong selection the effect can be considerable. For $4N_e s = 10$ and $\tau = 0.5$, the ML estimate was always infinity.

These simulations are a worst-case scenario, both in terms of the sample size and value of $N_e\mu$. With large data sets, in which the majority of sites are under weak selection, $4N_e s < 4$, as seems to be the case for selection on codon usage in Drosophila, estimates of the time since divergence should be reasonably unbiased. The strength of selection acting on codon usage is likely to be an underestimate. In the *mel-sim* comparison, the upper limit for the estimated strength of selection on codon usage in their MRCA is $\sim 4N_e s = 7.5$ (for the difference between the codons TTG and TTA for leucine in the gene *amyp*). Given our simulation results, it is possible that this may be half the true value, but it is unlikely to be an order of magnitude higher. We cannot rule out the possibility that a fraction of synonymous sites have much higher selection coefficients.