Substitution Rates in Drosophila Nuclear Genes: Implications for Translational Selection

Katherine A. Dunn, Joseph P. Bielawski and Ziheng Yang

Department of Biology, Galton Laboratory, University College, London NW1 2HE, United Kingdom

Manuscript received July 11, 2000
Accepted for publication October 12, 2000

ABSTRACT

The relationships between synonymous and nonsynonymous substitution rates and between synonymous rate and codon usage bias are important to our understanding of the roles of mutation and selection in the evolution of Drosophila genes. Previous studies used approximate estimation methods that ignore codon bias. In this study we reexamine those relationships using maximum-likelihood methods to estimate substitution rates, which accommodate the transition/transversion rate bias and codon usage bias. We compiled a sample of homologous DNA sequences at 83 nuclear loci from *Drosophila melanogaster* and at least one other species of Drosophila. Our analysis was consistent with previous studies in finding that synonymous rates were positively correlated with nonsynonymous rates. Our analysis differed from previous studies, however, in that synonymous rates were unrelated to codon bias. We therefore conducted a simulation study to investigate the differences between approaches. The results suggested that failure to properly account for multiple substitutions at the same site and for biased codon usage by approximate methods can lead to an artifactual correlation between synonymous rate and codon bias. Implications of the results for translational selection are discussed.

SYNONYMOUS substitutions do not affect the amino acid sequence of a protein, yet in some species of Drosophila, *Caenorhabditis elegans*, Arabidopsis, yeast, and enterobacteria, synonymous substitutions appear to be influenced by natural selection (Grosjean and Fiers 1982; Sharp *et al.* 1986; Sharp and Li 1987; Shields *et al.* 1988; Duret and Mouchiroud 1999). In Drosophila, synonymous codon usage is highly biased toward codons ending with G or C (Shields *et al.* 1988; Powell and Moriyama 1997; but see Rodríguez-Trelles *et al.* 1999), and there is evidence that selection favors substitutions to the preferred synonymous codons and limits substitutions to unpreferred synonymous codons (Akashi 1994, 1995, 1997). The exact causes of selection at synonymous sites, however, are less clear.

Selection for preferred codon usage in Drosophila has been suggested to enhance protein translation. Selective enhancement of translation is supported by the observation that the most frequent synonymous codons tend to match the most abundant tRNAs (Shields *et al.* 1988; Moriyama and Powell 1997b). In addition, a perceived negative correlation between silent divergence and codon bias (Shields *et al.* 1988; Sharp and Li 1989; Moriyama and Hartl 1993; Powell and Moriyama 1997) and a strong correlation between gene expression and codon bias (Shields *et al.* 1988; Powell and Moriyama 1997; Duret and Mouchiroud 1999) are consistent with greater selective pressure on the silent sites of genes with high codon bias. Finally, codon bias has been found to be lower in regions of low recombination (Kliman and Hey 1994; Comeron *et al.* 1999). This is expected if codon bias is maintained by selection, because in regions of low recombination the efficacy of selection is reduced due to linkage disequilibrium with other selected sites (Hill and Robertson 1966).

Although selective enhancement of translation appears to be the primary source of codon bias in Drosophila, it is less clear which of several mechanisms are operating. Translation could be enhanced by increasing the rate of elongation, reducing the cost of proofreading, increasing the accuracy of translation, or by any combination of those mechanisms (*e.g.*, Akashi and Eyre-Walker 1998). The majority of evidence supports the hypothesis that selection is acting to increase translational accuracy. Preferred codon usage is related to functional constraints, with the most conserved genes and most functionally important amino acid sites exhibiting the most biased codon usage (Akashi 1994). Furthermore, the synonymous substitution rate is perceived to be positively correlated with the nonsynonymous rate (Comeron and Kreitman 1998). These patterns indicate that synonymous substitutions are not independent of selective constraints acting on the amino acid composition of the protein. However, some data are inconsistent with the translational accuracy hypothesis. Codon usage bias is lower in longer genes (Comeron *et al.* 1999; Duret and Mouchiroud 1999), but selection to increase accuracy should be greater in longer genes

Corresponding author: Katherine A. Dunn, Department of Biology, University College, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: katherine.dunn@ucl.ac.uk

because the cost of proofreading is higher. In addition, codon bias increases within the first few hundred codons and then declines along Drosophila genes (Kliman and Eyre-Walker 1998). Hence the presence of additional selection pressures on codon bias, possibly to enhance translational efficiency (Comeron *et al.* 1999) or to maintain mRNA secondary structure (*e.g.*, Kirby *et al.* 1995), cannot be discounted.

Substitution rates in Drosophila genes contain important information about the effectiveness of selection. The relationship of synonymous substitution rate to codon bias, to nonsynonymous rate, and to location within a gene tells us something about the nature of selection (*e.g.*, Kliman and Eyre-Walker 1998). Previous studies of substitution rates employed approximate methods that neglect the effects of codon bias and apply *ad hoc* corrections for multiple substitutions. Recent studies suggest that approximate methods can lead to seriously biased estimates of synonymous and nonsynonymous substitution rates when codon usage is biased (Ina 1995; Yang and Nielsen 1998, 2000; Bielawski *et al.* 2000). An alternative is to use maximum likelihood (ML). The ML method is based on a Markov process model of codon substitution, which describes the changes between the sense codons in the genetic code and accounts for the transition/transversion rate ratio κ, different equilibrium codon frequencies, and the nonsynonymous/synonymous substitution rate ratio ω. Estimates of d_S and d_N are calculated according to their definitions from the maximum-likelihood estimates of model parameters (such as sequence divergence, κ, and ω). The probability theory corrects for unequal codon usage and for multiple substitutions in a straightforward manner. See Goldman and Yang (1994) and Yang and Nielsen (1998, 2000) for details.

The purpose of this study was to use ML methods to estimate substitution rates in Drosophila nuclear genes to investigate the relationship of synonymous rate to nonsynonymous rate and to codon usage bias. We compiled a sample of 83 loci with homologous DNA sequences from *Drosophila melanogaster* and at least one other species of Drosophila. Our analyses suggested a very different relationship between synonymous rate and codon usage bias as compared with all previous studies. A simulation study was thus conducted to investigate the source of this difference.

MATERIALS AND METHODS

Sequence data: Sequence data consist of 83 genes from *D. melanogaster* and at least one of the following species: *D. pseudoobscura, D. subobscura, D. simulans, D. yakuba,* and *D. virilis.* A list of these genes and their accession numbers is provided in the appendix. Some analyses were performed on a phylogeny of three species. Such an approach was possible for *D. melanogaster, D. pseudoobscura,* and *D. subobscura* (DmDpDsub) for 20 genes and for *D. melanogaster, D. simulans,* and *D. yakuba* (DmDsimDy) for 11 genes. Alternatively, larger sets of genes were analyzed in a pairwise fashion, *i.e., D. melanogaster vs. D. virilis* (DmDv; 39 genes), *D. melanogaster vs. D. pseudoobscura* (DmDp; 35 genes), and *D. melanogaster vs. D. simulans* (DmDsim; 24 genes).

Biased patterns of sequence evolution: G + C content at third codon positions (GC3) and codon usage bias, measured by the effective number of codons (ENC; Wright 1990), were computed for each gene, using the program Codon W of J. Penden. The assumption of homogeneous nucleotide frequencies among Drosophila species was tested using a chi-square test of a contingency table of nucleotide counts.

Estimation of the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site: Lineage-specific estimates of d_S and d_N were obtained from phylogenetic analyses of three species (DmDpDsub and DmDsimDy) using the ML method (Yang and Nielsen 1998). The assumption of a homogeneous d_N/d_S ratio among lineages was tested in both datasets by comparing two models. Model 0 assumed the same ratio for all three lineages of Drosophila, whereas model 1 allowed an independent d_N/d_S ratio for every lineage. Both models account for transition/transversion bias (κ) and unequal codon frequencies, which were determined using the empirical nucleotide frequencies at the three codon positions (F3 × 4 model; Yang 1999). Twice the log-likelihood difference between models 0 and 1 was compared with a χ^2 distribution with d.f. = (3 − 1) = 2.

Parameters d_S and d_N were estimated pairwise by ML for genes in the DmDv, DmDp, and DmDsim datasets (Goldman and Yang 1994). For comparison, we estimated d_S and d_N using two approximate methods: that of Nei and Gojobori (1986), referred to as NG, and that of Comeron (1995); the primary difference between them is that the method of Comeron (1995) corrects for transition/transversion rate bias while NG does not. The PAML package (Yang 1999) was used to implement NG and ML, while Comeron's program, K-Estimator 5.2, was used for Comeron's (1995) method.

Computer simulations: To understand the differences between methods, we simulated pairs of codon sequences using the evolver program of the PAML package (Yang 1999). Model parameters used were transition/transversion rate ratio (κ), d_N/d_S ratio (ω), codon frequencies (π_j), and sequence divergence (t). The κ value was set to one, and ω was the average from genes examined ($\omega = 0.06$). We used the empirically estimated codon frequencies from eight Drosophila genes with different ENC values: 53.4 from *sc*; 49.1 from *ade3*; 41.2 from *v*; 44.7 from *Gld*; 38 from *Gad1*; 33.7 from *Mlc1*; 32.3 from *Adh*; and 28.3 from *Amy-p*. Each set of codon frequencies was evaluated at a sequence divergence (t; measured by the expected number of nucleotide substitutions per codon) set to 0.4, 0.8, 1.2, 1.6, 2.0, 2.4, and 2.8. Values of t represent the range observed in the sampled Drosophila genes. In total, 56 pairs of sequences were simulated, each 1 million codons in length. d_S and d_N were estimated for each pair of sequences using both ML and NG methods. The method of Comeron (1995) was employed for a few parameter combinations only to examine the effect of transition/transversion rate bias.

RESULTS

Nucleotide (codon) frequencies in Drosophila nuclear genes: GC3 varied substantially among the 83 genes, ranging from 28 to 93%. This variation was characteristic of each dataset (DmDpDsub, 44–90%; DmDsimDy, 52–93%; DmDp, 36–90%; DmDv, 45–88%; and DmDsim, 28–91%). Chi-square tests suggested sig-

TABLE 1

Estimation of numbers of synonymous (d_S) and nonsynonymous substitutions (d_N) per site for each gene in the three-species datasets

				d_S			d_N		
Genes	κ	ω_0	$2\Delta\ell$	Lineage 1	Lineage 2	Lineage 3	Lineage 1	Lineage 2	Lineage 3
DmDpDsub									
Adh	1.39	0.0372	1.96	0.8512	0.2154	0.2816	0.0316	0.0080	0.0105
Adhr	1.43	0.0328	8.90*	1.5537	0.2061	0.1727	0.0337	0.0043	0.0157
Aprt	1.93	0.0526	0.10	1.2944	0.1629	0.2489	0.0681	0.0086	0.0131
ATPsyn b	1.89	0.0038	1.92	1.2969	0.1034	0.0994	0.0050	0.0004	0.0004
bcd	1.08	0.073	0.68	1.1001	0.0617	0.0939	0.0803	0.0045	0.0068
Cpy1	1.97	0.0237	0.54	0.8497	0.0911	0.0636	0.0202	0.0022	0.0015
Eno	1.99	0.0481	5.60	0.705	0.0944	0.0771	0.0339	0.0045	0.0037
Gad1	1.50	0.0136	0.86	0.9532	0.1933	0.1702	0.013	0.0026	0.0023
Gapdh2	1.97	0.0075	6.47*	1.6783	0.3632	0	0.0137	0	0.0015
Gld	1.37	0.0263	1.69	1.536	0.1869	0.1514	0.0404	0.0049	0.004
Gpdh	3.24	0.010	5.47	0.9283	0.1456	0.1031	0.0093	0.0015	0.001
Mlc1	0.90	0.0254	4.55	0.2602	0.0389	0.1289	0.0066	0.001	0.0033
ninaE	2.50	0.0194	2.40	0.5549	0.25	0.1486	0.0108	0.0049	0.0029
RpL32	5.14	0.0114	1.21	0.9485	0.0318	0.1548	0.0109	0.0004	0.0018
RpII215	2.19	0.0098	6.54*	1.2292	0.1939	0.1836	0.0142	0.0006	0.0012
ry	1.45	0.0317	3.25	1.5858	0.3762	0.3600	0.0503	0.0119	0.0114
sesB	1.55	0.0722	3.40	0.3327	0.0663	0.0625	0.0204	0.0048	0.0045
Sod	2.24	0.0293	2.68	2.1313	0.3591	0	0.0625	0.0105	0
Tpi	2.17	0.0173	3.63	1.8377	0.4563	0.1939	0.0318	0.0079	0.0034
Uro	1.31	0.0396	0.40	1.3219	0.3203	0.4429	0.0523	0.0127	0.0175
DmDsimDy									
ac	1.72	0.153	7.26*	0.0406	0.0747	0.1587	0.0088	0	0.0338
Adh	1.45	0.057	0.46	0.0467	0.0287	0.2079	0.0026	0.0016	0.0118
Amy-p	2.22	0.036	2.31	0.1928	0.0565	0.1956	0.0069	0.0020	0.007
GstD1	3.01	0.057	11.28*	0.0877	0.0134	0.0641	0.0059	0.1797	0
l(1)sc	1.63	0.060	4.74	0.0380	0.0245	0.2814	0.0023	0.0015	0.0168
nullo	1.97	0.093	4.37	0.0735	0.0871	0.6073	0.0068	0.0081	0.0562
per	3.17	0.026	10.55*	0.1363	0.0014	0.1219	0	0.5328	0.0198
Pgi	3.87	0.024	2.97	0.0559	0.0292	0.2554	0.0014	0.0007	0.0062
sc	1.23	0.100	1.80	0.0754	0.0373	0.1875	0.0075	0.0037	0.0167
Sry-beta	2.70	0.088	0.20	0.1176	0.0179	0.1998	0.0103	0.0016	0.0175
Zw	2.51	0.061	1.87	0.0726	0.1712	0.2004	0.0044	0.0104	0.0122

Gene symbols are from FlyBase (1999). DmDpDsub refers to the *Drosophila melanogaster*, *D. pseudoobscura*, *D. subobscura* dataset, and lineage 1, lineage 2, and lineage 3 indicate the first, second, and third species of this dataset, respectively. DmDsimDy refers to the *D. melanogaster*, *D. simulans*, *D. yakuba* dataset and lineage 1, lineage 2, and lineage 3 indicate the first, second, and third species of the dataset, respectively. Model 0 assumed one d_N/d_S ratio for all branches and model 1 assumed brach-specific d_N/d_S ratios. $2\Delta\ell$ is twice the difference in log-likelihood scores for models 0 and 1. Values of d_S and d_N were those obtained from the best-fit model (model 0 or model 1). *, a significant difference between the two models.

nificant heterogeneity in nucleotide composition among lineages for 4 of 20 genes in DmDpDsub (*Gapdh2*, *Gpdh*, *RpII215*, and *ry*), 8 of 35 genes in DmDp [*Gapdh2*, *Gpdh*, *l(2)gl*, *Rh2*, *RpII215*, *ry*, *Tl*, and *Ubx*], and 11 of 39 genes in DmDv [*Adh*, *Amy-p*, *Cdc37*, *fu*, *gbb*, *Gpdh*, *Kr*, *l(2)tid*, *lama*, *nos*, and *Sry-beta*].

Consistent with patterns of nucleotide bias, codon usage varied greatly among genes; the ENC values ranged from extreme bias at 26.7 (*GstD1*) to no bias at 61 (*sc*). Each dataset exhibited substantial variation in codon bias among genes (DmDpDsub, 27.3–59.8; DmDsimDy, 26.7–61; DmDp, 28.6–56.6; DmDv, 29.2–61; and DmDsim, 26.7–61). Codon bias was not related to heterogeneity in nucleotide composition among lineages. For example, genes with homogeneous nucleotide composition among lineages exhibited both highly biased (*e.g.*, gene *Eno*, ENC = 28.6) and unbiased (*e.g.*, gene *ac*, ENC = 56.1) codon usage.

Lineage-specific patterns of synonymous and nonsynonymous substitution: ML estimation of d_N and d_S was carried out for each gene in the DmDpDsub and DmDsimDy datasets (Table 1). Constancy of nonsynonymous/synonymous rate ratios (d_N/d_S) was tested, and d_N/d_S ratios were largely homogeneous over lineages (Table 1). Homogeneity was rejected in only 3 of the 20 genes in DmDpDsub and 3 of the 11 genes in DmDsimDy. This trend differs from that observed in mammals, where over half of genes surveyed exhibited

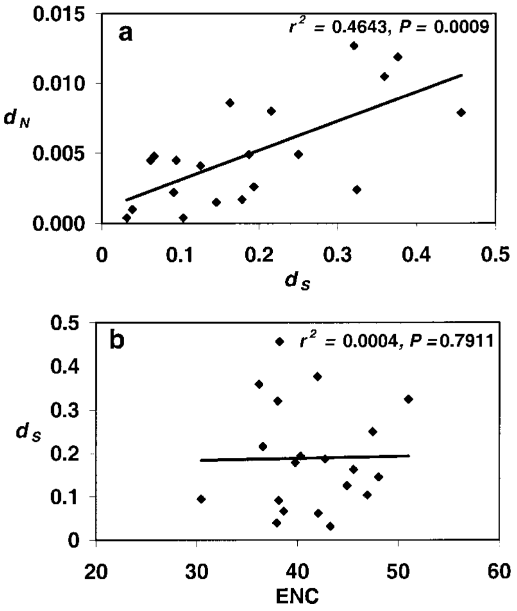

Figure 1.—The relationship between d_N and d_S (a) and between d_S and ENC (b) for *Drosophila pseudoobscura*. Estimates of d_S and d_N were obtained from lineage-specific analyses.

significant variation in selective constraints among lineages (Yang and Nielsen 1998; Bielawski *et al.* 2000).

Genes exhibiting high synonymous or nonsynonymous rates in one lineage tended to exhibit high rates in other lineages as well. For example, the coefficient of determination (r^2) of d_S estimates between Dm and Dp is r^2 = 0.5098 (P = 0.0004), and the coefficient of determination of d_N is r^2 = 0.5005 (P = 0.0005). Similar patterns were reported for mammalian genes (Bulmer *et al.* 1991; Mouchiroud *et al.* 1995). Substantial variation in d_N/d_S was found among genes, as they are under very different selective constraints (Table 1).

The relationship between synonymous and nonsynonymous substitution rates was evaluated by linear correlation using rate estimates of Table 1 for five species. Estimates of d_S were positively correlated with d_N in every lineage (Dp, r^2 = 0.4643, P = 0.0009; Dsub, r^2 = 0.7312, $P \ll 0.0001$; Dsim, r^2 = 0.5722, P = 0.0071; Dy, r^2 = 0.3945, P = 0.0385; Dm, r^2 = 0.4587, $P \ll 0.0001$). The plot for *D. pseudoobscura* is presented as an example (Figure 1a). These results are consistent with a previous analysis of substitution rates in Drosophila (Comeron and Kreitman 1998). However, this study represents a wider sample of Drosophila species and the correlations reported here are stronger than previously reported.

Synonymous substitution rates were independent of codon bias, as evaluated by linear regression of d_S and ENC, in every lineage (Dp, r^2 = 0.0004, P = 0.7911; Dsub, r^2 = 0.0138, P = 0.6218; Dsim, r^2 = 0.151, P = 0.2376; Dy, r^2 = 0.0045, P = 0.8446; Dm, r^2 = 0.0079, P = 0.6405). The plot for *D. pseudoobscura* is presented as an example (Figure 1b). All previous analyses differ from ours in suggesting synonymous substitution rates are negatively correlated with codon bias (Shields *et al.* 1988; Sharp and Li 1989; Moriyama and Hartl 1993; Powell and Moriyama 1997). Previous studies, however, also differ from ours in using approximate estimation methods, which ignore codon usage bias. Furthermore, in this study we estimated d_S per lineage rather than between pairs of sequences. In the next section, we explore what factors may be responsible for this conflict.

Reconciling differences between ML and approximate methods: Models employed in ML estimation of substitution rates explicitly assumed that nucleotide/codon frequencies were at equilibrium (Goldman and Yang 1994). To see whether violation of this assumption influenced our correlation analysis, we reevaluated the relationship between d_S and ENC using only stationary genes. To keep more genes in the dataset and to facilitate direct comparison with approximate methods, we apply the ML method to two-taxa datasets. This pairwise ML analysis led to the same conclusion as the phylogeny-based ML analyses; *i.e.*, synonymous rates were independent of codon bias (Figure 2). However, when either NG or Comeron's (1995) method was used, d_S was negatively correlated with ENC for both the DmDv (NG, r^2 = 0.5036, $P \ll 0.0001$; Comeron, r^2 = 0.4615, P = 0.0001) and DmDp (NG, r^2 = 0.6754, $P \ll 0.0001$; Comeron, r^2 = 0.6127, $P \ll 0.0001$) datasets (Figure 2), a result consistent with previous studies (Sharp and Li 1989; Comeron and Aguadé 1996; Powell and Moriyama 1997). However, in the DmDsim dataset, the approximate methods lead to the same conclusion as the ML method; d_S was independent of ENC (NG, r^2 = 0.0995, P = 0.1332; Comeron, r^2 = 0.0583, P = 0.2557; Figure 2). Because results were similar for the two approximate methods, only results for NG are presented in Figure 2. On the basis of these results we conclude that the conflict between approximate methods and ML cannot be attributed to use of a pairwise approach, the sample of genes, or nonstationarity of nucleotide frequencies.

To examine the effect of transition/transversion bias and codon bias on estimation of d_S, we changed parameters of the codon model to reanalyze the two-taxa datasets by ML. The effect of ignoring transition/transversion bias was evaluated by setting $\kappa = 1$, but allowing for biased codon usage. Results obtained using this codon model did not differ from those obtained using the full codon model; *i.e.*, there was no significant relationship between d_S and ENC for DmDv (r^2 = 0.140, P = 0.060) and DmDp (r^2 = 0.051, P = 0.256). The effect of ignoring codon bias was evaluated by assuming equal codon frequencies, but allowing for transition/transversion bias. This codon model produced results different from those obtained previously and matched the approximate methods, with a significant correlation between d_S and ENC in DmDv (r^2 = 0.435, P = 0.0005) and DmDp (r^2 = 0.6281, $P \ll 0.0001$). These findings suggest that the different treatment of codon bias was responsible for the conflict between ML and approximate methods in the relationship between d_S and ENC.

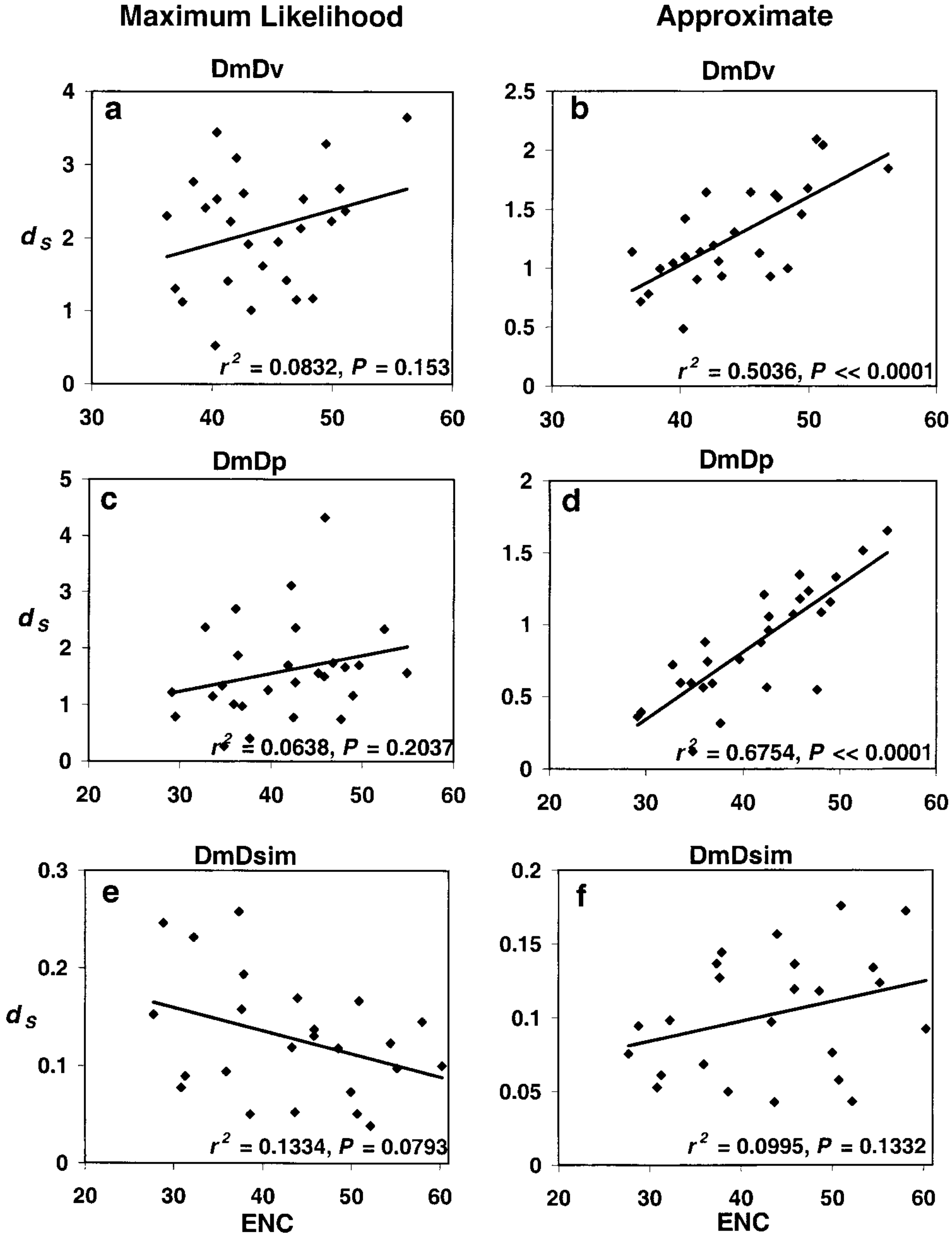

FIGURE 2.—The relationship between pairwise estimates of d_S and mean ENC (a–f). Pairwise estimates of d_S were computed using both ML (GOLDMAN and YANG 1994) and NG (NEI and GOJOBORI 1986). DmDv indicates pairwise comparisons between *D. melanogaster* and *D. virilis*; DmDp, pairwise comparisons between *D. melanogaster* and *D. pseudoobscura*; and DmDsim, pairwise comparisons between *D. melanogaster* and *D. simulans*.

However, in one pairwise comparison (DmDsim), approximate and ML methods were in agreement, yet codon usage in this dataset (mean ENC = 43) was just as biased as in the other two sets of genes (DmDv, mean ENC = 44; DmDp, mean ENC = 41) where approximate and ML methods differed. Hence, there must have been an additional factor. In both the DmDv and DmDp datasets, where approximate methods produced a positive relationship between d_S and ENC, uncorrected percent sequence divergences (p) were very large (DmDv, mean p = 21.6 ± 6.0%; DmDp, mean p = 16.3 ± 5.8%). In the DmDsim dataset, where d_S was independent of ENC, divergences were very low (mean p = 3.2 ± 1.4%). This pattern suggested a possible "saturation" effect. Consequently, we hypothesized that differences between approximate and ML methods might be related to the combined effects of sequence divergence and codon usage bias.

Simulation studies: Simulations were performed to evaluate the hypothesis that improper treatment of both divergent sequences and codon usage bias by approximate methods could have led to a significant positive correlation between d_S and ENC when, in reality, they were independent. Codon sequences were simulated under values of t chosen to reflect the range exhibited among Drosophila genes, and codon frequencies were modeled after observed frequencies of Drosophila genes (see MATERIALS AND METHODS). Simulated codon sequences were analyzed using both NG and ML (F3 × 4), and these results were compared with the true values employed in the simulation (Figure 3).

Both sequence divergence and codon bias had an

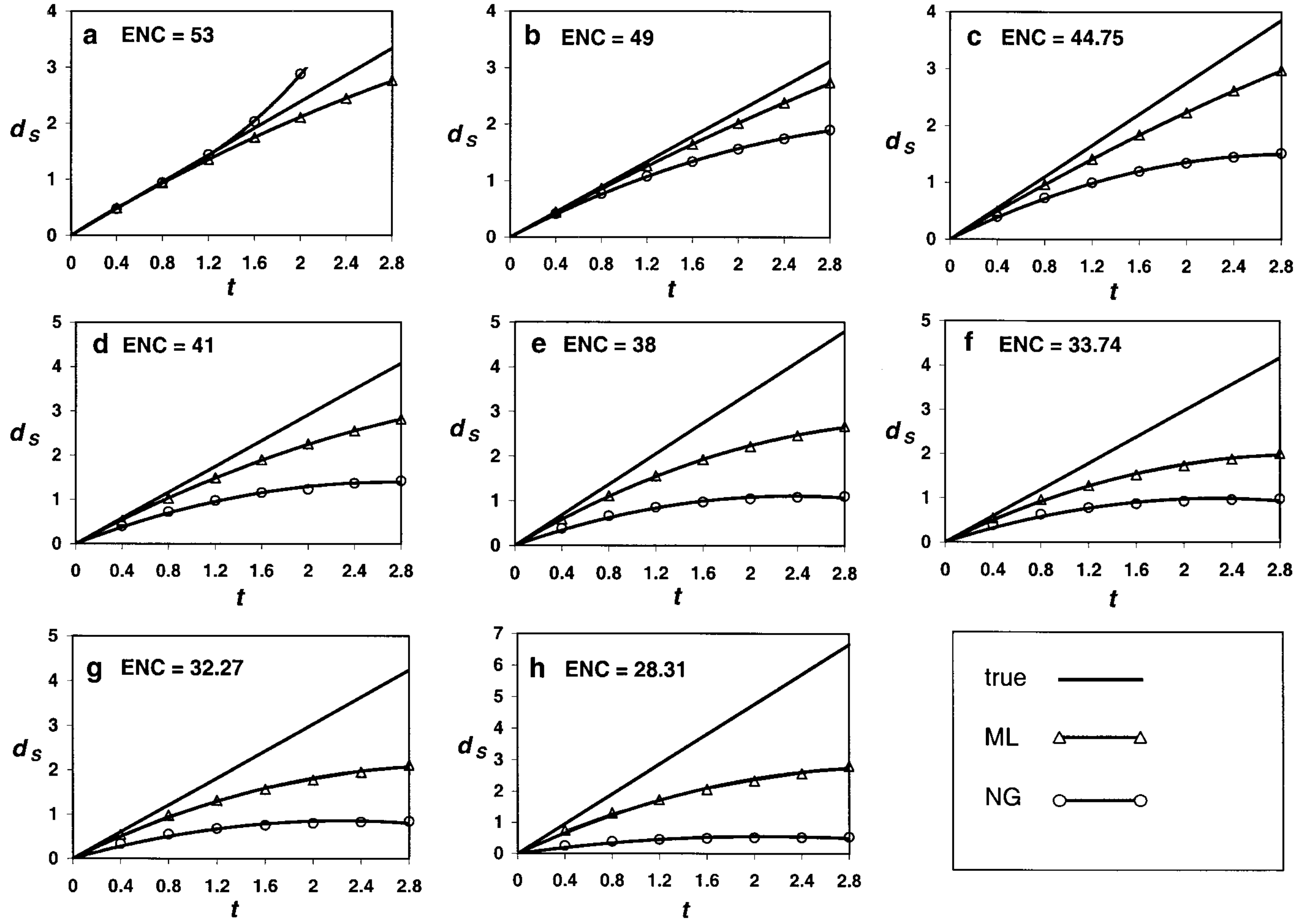

FIGURE 3.—Estimates of d_S in simulated data by NG (○) and ML (F3 × 4; △) plotted against sequence divergence t (the expected number of nucleotide substitutions per codon). Data for a pair of sequences, each of 1 million codons, were simulated for different codon bias measured by ENC.

effect on estimates of d_S (Figure 3). Only at low sequence divergences ($t < 1.6$) and relatively unbiased codon usage (ENC = 53) did both NG and ML (F3 × 4) produce unbiased estimation of d_S. With one exception (Figure 3a), d_S was underestimated by both methods with increasing levels of sequence divergence (Figure 3, b–h), and the NG method involved a much more serious estimation bias than ML. Because equilibrium codon frequencies obtained using the F3 × 4 model (9 frequency parameters) do not perfectly match the empirical Drosophila frequencies, some degree of error was expected for ML. Use of the more parameter-rich model (60 frequency parameters) produced estimates of d_S essentially identical to the true values (data not shown). Differences between the analysis of real data using the NG and ML (F3 × 4) methods are consistent with differences observed in our simulation study; *i.e.*, ML estimates of d_S were larger than estimates obtained using NG. These simulations further suggested that the ML (F3 × 4) estimates of d_S for Drosophila genes are, themselves, likely to be underestimates of the true values. Although the F3 × 4 model is biased in cases of extreme codon bias, this model was acceptable over a wide range of codon biases and consistently outperformed the NG method.

Plots of d_S (Figure 3) are reminiscent of the "saturation effect" on plots of uncorrected sequence divergence. For d_S, however, the "ceiling" appears to be related to levels of codon usage bias; the effect is relatively minor when codon usage is unbiased and extreme when codon usage is highly biased (Figure 3). For example, in the most extreme case of codon bias (ENC = 28), the estimate of d_S obtained using the NG method peaked at 0.5 ($t > 0.8$), although true values of d_S range from 0.95 when $t = 0.4$ to 6.6 when $t = 2.8$. The only exception to the general pattern occurred when codon usage was relatively unbiased (Figure 3a). In this case, when $1.2 < t < 2.0$, NG overestimated rather than underestimated d_S, and when $t > 2.0$ the method yielded invalid estimates of d_S. Interestingly, MUSE (1996) suggested that even when there is no codon usage bias, *ad hoc* corrections for multiple substitutions applied in the approximate methods could introduce substantial bias.

The dramatic effect of codon usage bias on estimation

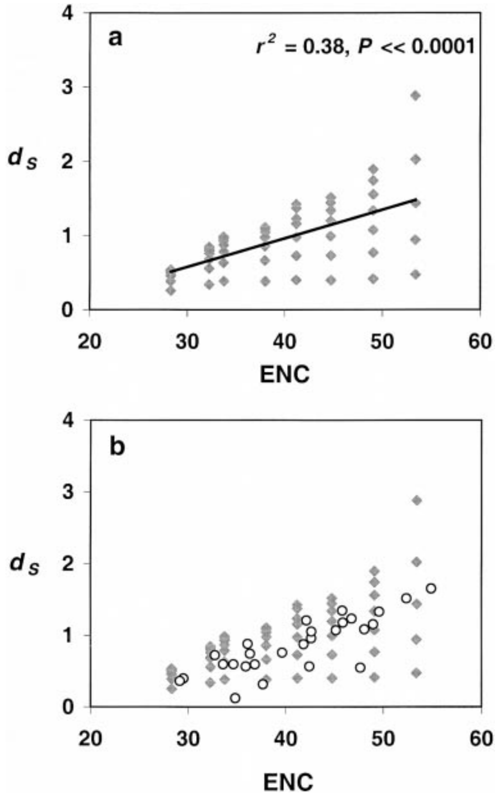

Figure 4.—(a) Approximate estimates of d_S by NG obtained in the simulation (Figure 3) plotted against ENC. Multiple estimates of d_S for each ENC correspond to the seven values of t. Note the true d_S is independent of ENC, but a positive correlation is suggested by NG. (b) The same data as in a but with NG estimates from the pairwise comparisons between *Drosophila melanogaster* and *D. pseudoobscura* superimposed (○).

of d_S is caused mainly by its effect on counting of synonymous (S) and nonsynonymous (N) sites. When codon usage is biased but ignored (Figure 3, b–h), the number of synonymous sites (S) was overestimated, leading to an underestimation of d_S. For example, in the most extreme case (ENC = 28, t = 2.8), most codons end with C or G and most changes at the third position are transversions between C and G, which are more likely to be nonsynonymous than random changes between nucleotides. As a result, the proportion of synonymous sites is as low as S = 8.53%. The NG method assumes unbiased codon usage or equal rates of change between nucleotides and expects more frequent transitional or synonymous changes, giving S = 23.6%, with almost a fourfold difference. When there is little codon bias (ENC = 53), NG reliably estimated S (NG, S = 23.4%; true value, S = 23.3%). Thus the underestimation of d_S by NG is more serious for more extreme codon bias (Figure 3). In Figure 4a, the NG estimates of d_S corresponding to different sequence divergences (t) from the simulation (Figure 3) were plotted against ENC. Although d_S and ENC are independent, NG incorrectly indicated a positive correlation, due to the underestimation of d_S at high divergences and extreme codon bias.

The ML method (not plotted) indicated no correlation between d_S and ENC (r^2 = 0.00006, P = 0.95). The pattern seen in the simulated data is consistent with that in the real data. This is most clear when NG estimates of d_S from the real data (DmDp, Figure 2d) were superimposed onto the approximate estimates from the simulations (Figure 4b). These data suggest that the significantly positive correlation between d_S and ENC indicated by the approximate methods is an artifact of these methods' failure to properly account for the combined effects of codon bias and sequence divergence.

DISCUSSION

Codon usage in Drosophila varies considerably between genes and does not appear random. Highly conserved genes, and functionally important sites, exhibit highly biased synonymous codon usage (Akashi 1994). Akashi (1994) theorized that these patterns resulted from a balance between selection for translational accuracy, mutation, and drift in finite populations. On the basis of this model, he predicted that synonymous substitution rates should be positively correlated with nonsynonymous substitution rates. However, he did not find a significant correlation. More recently, Comeron and Kreitman (1998) reported that when data from three species of Drosophila are combined, synonymous rates are positively correlated with nonsynonymous substitution rates. Our findings confirm those of Comeron and Kreitman (1998) and further indicate that this correlation is a feature of all five species of Drosophila. In addition, our analysis implies that the correlation between synonymous and nonsynonymous rates is stronger than previously thought, suggesting that synonymous substitution rates in these Drosophila species are not independent of selective constraints acting at the amino acid level. Consistent with patterns of codon usage (Akashi 1994), these data support the hypothesis that codon bias in Drosophila is influenced by selection for translational accuracy.

However, our findings differ from all previous studies (Shields *et al.* 1988; Sharp and Li 1989; Moriyama and Gojobori 1992; Moriyama and Hartl 1993; Carulli *et al.* 1993; Powell and Moriyama 1997) in suggesting that synonymous substitution rates are not correlated with codon bias. Through ML analysis under different models as well as computer simulation, we show that previous reports of a significant correlation between d_S and ENC might be an artifact of inadequate correction for the combined effects of saturation and biased codon usage. However, Powell and Moriyama (1997) suggested that artifacts of inadequate correction for codon bias could not be responsible for a correlation between d_S and ENC because in their study they employed a method (Moriyama and Powell 1997a) that was claimed to correct for codon bias. However, this method corrects for base composition bias only when correcting for multiple substitutions and assumed unbiased codon usage in the important steps of counting sites and differences (Yang and Nielsen 2000). Because the major source of bias in d_S and d_N arises from biased estimates of S and N (Yang and Nielsen 2000; see also above), the method of Moriyama and Powell (1997a) does not correct effectively for codon usage bias.

The effect of saturation is still evident in the analysis of Powell and Moriyama (1997). They observed a very strong correlation between d_S and codon usage bias in distantly related species pairs, but a much weaker correlation in a more recently diverged pair of species, *D. melanogaster* and *D. simulans.* Powell and Moriyama (1997) dismissed the weak correlation as an effect of sampling errors at small sequence divergences and possible shared polymorphisms. Our simulation study suggests an opposite conclusion: that is, the weak correlation is closer to the true relationship, while the strong correlation is an artifact; *i.e.*, because *D. melanogaster* and *D. simulans* have a more recent divergence, this species pair provided the least biased estimate of the relationship between d_S and codon bias when approximate methods are used.

Our analysis suggests that the discrepancy between approximate and ML methods was caused mainly by codon usage bias and not by the transition/transversion bias (see also Yang and Nielsen 1998, 2000; Bielawski *et al.* 2000). The approximate methods of Nei and Gojobori (1986) and Comeron (1995) led to the same conclusion. The ML method both with and without accounting for the transition/transversion bias led to the same conclusion, which is different from that of the approximate methods. To test this notion further, a small simulation study was conducted using the method of Comeron (1995), which corrects for transition/transversion bias. The transition/transversion ratio was fixed at $\kappa = 2$, and three levels of codon bias (ENC = 49, 38, and 28.31) were used. Simulated sequences comprised only 30,000 codons as the program of Comeron did not handle very long sequences. Results, shown in Figure 5, indicate that simply accounting for the transition/transversion bias (*i.e.*, the method of Comeron 1995) does not reduce the bias in estimates of synonymous substitution rates. In fact, at moderate codon bias (ENC = 49), NG is closer to the real value of d_S than the method of Comeron (1995). This is because the bias introduced by ignoring codon usage is in the opposite direction to that produced by ignoring the transition/transversion ratio, and the two biases partially cancel out in NG. This pattern was discussed by Yang and Nielsen (2000). However, with greater codon bias (*i.e.*, ENC > 49), the effect of ignoring codon bias becomes much larger than ignoring transition bias, hence both NG and Comeron's (1995) methods yield nearly identical estimates of d_S.

Given the evidence for selection for translational accuracy in Drosophila, our finding of lack of correlation between synonymous rate and codon usage bias is puzzling. Models of translational selection predict that selection against substitutions to unpreferred codons at functionally important amino acid sites will result in a positive correlation between d_S and ENC (Shields *et al.* 1988; Sharp and Li 1989; Moriyama and Hartl 1993; Powell and Moriyama 1997). However, this prediction assumes that mutation pressure does not contribute to codon bias. Kliman and Hey (1994) found a correlation between base composition of third codon positions and introns, and they suggested that at least 10% of variation in codon bias can be explained by mutation pressure.

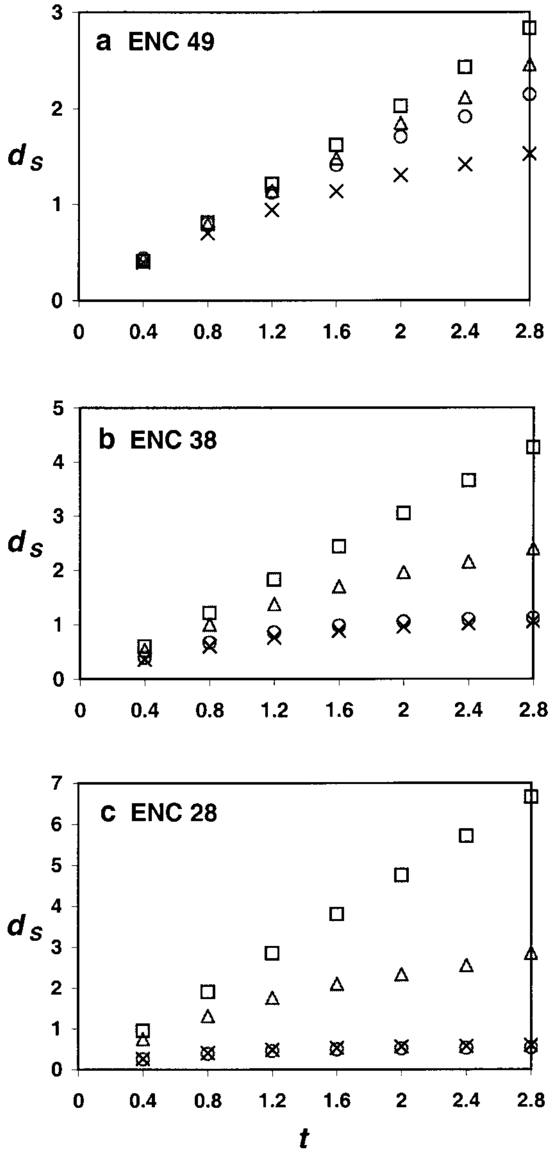

Figure 5.—Estimates of d_S in simulated data by NG (○), ML (F3 × 4; △), and Comeron (1995) (×) plotted against sequence divergence t (the expected number of nucleotide substitutions per codon). (□) Real values of d_S. Data for a pair of sequences, each of 30,000 codons, were simulated under $\kappa = 2$ and three different levels of codon bias (a–c).

Variable mutation pressures can lead to a negative, rather than positive, correlation between d_S and ENC (Bielawski *et al.* 2000). If mutation pressure produces a negative relationship between d_S and ENC in Drosophila, it might confound the positive relationship between d_S and ENC expected under translational accuracy.

Comeron and Kreitman (1998) found that, among the majority of codons having double nucleotide substitutions in Drosophila, synonymous substitutions represented a shift to an unpreferred rather than a preferred codon, a pattern inconsistent with positive codon selection. Comeron and Kreitman (1998) hypothesized that the explanation for this pattern was a relaxation of selective constraints at these sites. They suggested a covarion-like model whereby the effect of selection on individual codons, for translational accuracy, changes over time. In this model, any relaxation of constraints at an amino acid site also includes relaxation of constraints on synonymous codon usage at the same site. Fluctuations in selective constraints could obscure the relationship between d_S and ENC predicted by previous models of translational selection because d_S is measured over time and ENC is measured at one point in time. In genes having a large proportion of sites where selective constraints have changed over time, substitution rates will be higher as compared to genes having a smaller fraction of sites subject to changing selective pressures. At any one point in time these genes could have similar values of ENC. Interestingly, we observed considerable variation in d_S among genes with similar values of ENC. Moreover, the covarion-like model of Comeron and Kreitman (1998) predicts a positive correlation between synonymous and nonsynonymous substitution rates, which we also observed among Drosophila.

Whatever the causes of codon bias in Drosophila, the results of this study, and others (Kliman and Hey 1994; Comeron and Kreitman 1998; Kliman and Eyre-Walker 1998; Comeron *et al.* 1999; Duret and Mouchiroud 1999; McVean and Vieira 1999), indicate that the origins of codon bias in Drosophila are more complex than previously thought. Unbiased estimation of substitution rates will be a crucial aspect of unraveling the origins of codon bias in Drosophila.

We are grateful for the comments of two anonymous reviewers. This study was supported by a Biotechnology and Biological Sciences Research Council grant (31/G10434) to Z.Y.

LITERATURE CITED

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

Akashi, H., 1997 Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. Genetics **147:** 1989–1991.

Akashi, H., and A. Eyre-Walker, 1998 Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8:** 688–693.

Bielawski, J., K. A. Dunn and Z. Yang, 2000 Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics **156:** 1299–1308.

Bulmer, M., K. H. Wolfe and P. M. Sharp, 1991 Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88:** 5974–5978.

Carulli, J. P., D. E. Krane, D. L. Hartl and H. Ochman, 1993 Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics **134:** 837–845.

Comeron, J. M., 1995 A method for estimating the number of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41:** 1152–1159.

Comeron, J. M., and M. Aguadé, 1996 Synonymous substitutions in the *Xdh* gene of Drosophila: heterogeneous distribution along the coding region. Genetics **144:** 1053–1062.

Comeron, J. M., and M. Kreitman, 1998 The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics **150:** 767–775.

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

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.

FlyBase, 1999 The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res. **27:** 85–88. (Available from http://flybase.bio.indiana.edu/.)

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

Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Ina, Y., 1995 New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40:** 190–226.

Kirby, D. A., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Kliman, R. M., and A. Eyre-Walker, 1998 Patterns of base composition within the genes of *Drosophila melanogaster*. J. Mol. Evol. **46:** 534–541.

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.

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 D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

Moriyama, E. N., and J. R. Powell, 1997a Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J. Mol. Evol. **45:** 378–391.

Moriyama, E. N., and J. R. Powell, 1997b Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45:** 514–523.

Mouchiroud, D., C. Gautier and G. Bernardi, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40:** 107–113.

Muse, S. V., 1996 Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13:** 105–114.

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

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

Rodríguez-Trelles, F., R. Tarrío and F. J. Ayala, 1999 Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics **153:** 339–350.

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., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

Sharp, P. M., T. M. F. Tuohy and K. R. Mosurski, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5139.

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.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Yang, Z., 1999 *Phylogenetic Analysis by Maximum Likelihood (PAML), Version 2.* University College, London.

Yang, Z., and R. Nielsen, 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46:** 409–418.

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

Communicating editor: J. Hey

APPENDIX

Drosophila genes and accession numbers

Genes	*Drosophila melanogaster*	*D. pseudoobscura*	*D. subobscura*	*D. simulans*	*D. yakuba*	*D. virilis*
ac	M17120			X62400	AB005751	
Acp26Ab	X70888			X70899		
Act88F	M18826			M87274		
ade3	X06286	X06285				
Adh	X78384	X62181	M55545	X00607	X54120	U26846
Adhr	X78384	Y00602	M55545			
Amy-p	L22721	X76241		D17734	D17738	U02029
Amyrel	U69607	U82556				
Aprt	M18432	L06281	AF025800			
ATPsyn-b	X71013	AF025802	AF025801			X86017
bcd	X07870	X55735	X78058			
boss	L08133					L08132
bw	L23543					L37035
Cdc37	L32839					L37055
Cpy1	M62398	AF025803	AF025804			
csw	M94730					U22356
Cyt-b5	X15008					U12418
Ddx1	U34773					U34779
dpp	U63857	U63856		U63854		
e(r)	L36921					U66868
elav	M21152					M61748
en	M10017					X04727
Eno	X17034	AF025805	AF025806			
Est-6	M33780			L10670		
Est-P	M33780	M55907				
exu	S72363	L22554				
fu	X80468					U20586
Gad1	X76198	AF025807	AF025808			
Gapdh2	M11255	AF025809	AF025810			
gbb	M84795					U48595
gl	X15400					U39746
Gld	M29298	M29299	AF025811	U63324		
Gpdh	X67650	U47885	U47872	AF085163		D10697
GstD1	X14233			M84577	M84580	
h	X15904					M87885
hb	Y00274					X15359
His1	X04073					L76558
Hsp83	X03810	X03812				X03813
kni	X13331					L36177
Kr	X03414					U49856

(*continued*)

APPENDIX

(Continued)

Genes	*Drosophila melanogaster*	*D. pseudoobscura*	*D. subobscura*	*D. simulans*	*D. yakuba*	*D. virilis*
ksr	U43583					U43584
l(1)sc	X71806			AB005802	AB005800	
l(2)gl	M17022	X73259				
l(2)tid	X95241					Y07700
lama	U57314					U57315
mam	X54251					M92914
Mlc1	M10125	L08052	AF025812	L08051		L08053
ninaE	K02315	X65877	AF025813			
nos	M72421					U24695
nullo	X65444			U44733	U44732	
osk	M63492					L22556
Pcp	X06286	X06285				
pdm2	M93149					U14723
per	AF033029			L07829	X61127	
Pgd	M80598			U02288		
Pgi	L27555			L27552	L27673	
ref(2)P	X16993			U23930		
Rh2	M12896	X65878				
Rh3	M17718	X65879				
Rh4	AH001040	X65880-1				M77281
RpII215	M27431	Y18876	Y18879			
RpL32	Y13939	S59382	M21333			
run	X56432	U22357				U22358
ry	Y00307	M33977	Y08237			
sala	X57474			M21227		
sc	M17119			AB005801	AB005799	
sesB	S43651	AF025798	AF025799			
sina	M38384					M77282
Sod	M24421	U47871	U47888	X15685		X13831
Sry-alpha	X03121			U64718	U64719	
Sry-beta	X03121					AF084637
su(Hw)	Y00228					Z25520
Su(var)	M57574					M88753
Tl	M19969	L25390				
Tpi	X57576	AF025814	AF025815			
tra	M17478			X66930		
Ubx	U31961	X05179				
Uro	X51940	X57113	AF025816			
v	M34147			U23204		
w	X51749			U64875		
Yp1	V00248					U52124
z	Y00049					M76700
Zw	L13880			L13894	U42750	

Abbreviations for genes follow FlyBase (1999).