Inferring Weak Selection From Patterns of Polymorphism and Divergence at "Silent" Sites in Drosophila DNA

Hiroshi Akashi

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received June 29, 1994
Accepted for publication October 19, 1994

ABSTRACT

Patterns of codon usage and "silent" DNA divergence suggest that natural selection discriminates among synonymous codons in Drosophila. "Preferred" codons are consistently found in higher frequencies within their synonymous families in *Drosophila melanogaster* genes. This suggests a simple model of silent DNA evolution where natural selection favors mutations from unpreferred to preferred codons (preferred changes). Changes in the opposite direction, from preferred to unpreferred synonymous codons (unpreferred changes), are selected against. Here, selection on synonymous DNA mutations is investigated by comparing the evolutionary dynamics of these two categories of silent DNA changes. Sequences from outgroups are used to determine the direction of synonymous DNA changes within and between *D. melanogaster* and *Drosophila simulans* for five genes. Population genetics theory shows that differences in the fitness effect of mutations can be inferred from the comparison of ratios of polymorphism to divergence. Unpreferred changes show a significantly higher ratio of polymorphism to divergence than preferred changes in the *D. simulans* lineage, confirming the action of selection at silent sites. An excess of unpreferred fixations in 28 genes suggests a relaxation of selection on synonymous mutations in *D. melanogaster*. Estimates of selection coefficients for synonymous mutations ($3.6 < |N_e s| < 1.3$) in *D. simulans* are consistent with the reduced efficacy of natural selection ($|N_e s| < 1$) in the three- to sixfold smaller effective population size of *D. melanogaster*. Synonymous DNA changes appear to be a prevalent class of weakly selected mutations in Drosophila.

THE strongest evidence for natural selection at "silent" sites in DNA comes from studies of codon usage in *Escherichia coli* and *Saccharomyces cerevisiae*. In these species, synonymous codon usage is biased toward a subset of preferred codons, which generally code for the most abundant tRNA(s) for each amino acid (IKEMURA 1981; BENNETZEN and HALL 1982; GROSJEAN and FREIRS 1982). The degree to which codon usage is biased varies considerably between genes and correlates strongly with gene expression levels (IKEMURA 1985). Silent DNA divergence between *E. coli* and *Salmonella typhemurium* is inversely related to codon usage bias, which suggests the action of purifying selection at preferred codons (SHARP and LI 1987). These patterns, termed "major codon preference," are consistent with natural selection discriminating among synonymous codons to enhance translational efficiency and/or accuracy (BULMER 1991).

Patterns of codon usage and silent DNA evolution in *D. melanogaster* are similar to those found in *E. coli* and yeast (SHIELDS *et al.* 1988). Codon usage is biased toward a subset of (mostly G/C-ending) synonymous codons for each amino acid. Relative tRNA abundances have been quantified for only a few amino acids, and, where the data exist, codon preferences correlate with the abundant tRNA (see SHIELDS *et al.* 1988). Although relative expression levels are difficult to quantify in multicellular organisms, where gene expression can be tissue- and developmental-stage specific, highly expressed loci such as ribosomal protein genes and *Adh* show highly biased codon usage, whereas less highly expressed genes such as *Adhr* show more equal usage among synonymous codons. Silent DNA divergence between Drosophila species is inversely related to codon usage bias, which again suggests varying levels of purifying selection at different loci (SHARP and LI 1989; CARULLI *et al.* 1993).

Two observations add further support for the role of selection at silent sites in Drosophila. Evidence for lower codon usage bias in regions of low recombination in the *D. melanogaster* genome is consistent with theoretical predictions of the reduced efficacy of selection in such regions (KLIMAN and HEY 1993a). An association between codon bias and functional constraint at the protein level has been found in Drosophila, suggesting that natural selection discriminates among synonymous codons to enhance the accuracy of protein synthesis (AKASHI 1994).

Major codon preference can be modeled under a simple form of selection-mutation-drift at silent sites (SHARP and LI 1986; BULMER 1991). Selection favors a subset of synonymous codons for each amino acid,

Author e-mail: akas@midway.uchicago.edu

whereas mutation pressure and genetic drift allow nonpreferred codons to persist. Gene expression levels, protein functional constraint and possibly other factors determine the intensity of selection at silent sites in a given gene. Under this model, major codon preference is the result of both positive selection for mutations that increase the frequency of preferred codons (preferred changes) and selection against deleterious mutations in the opposite direction from preferred to nonpreferred codons (unpreferred changes). Although the strength of selection may vary among genes, positive and negative selection will differentiate the dynamics of these classes of silent changes across loci showing codon bias.

Relatively high levels of synonymous polymorphism in Drosophila (KREITMAN 1983; AGUADÉ *et al.* 1992; AYALA and HARTL 1993; KLIMAN and HEY 1993b) suggest that selection at silent sites may be very weak. Theory has shown, however, that the *ratio* of levels of polymorphism within species to divergence between species is sensitive to very small differences in the fitness effect of mutations (KIMURA 1983). Subdividing synonymous DNA changes into putatively favorable and deleterious mutations provides an opportunity to investigate natural selection by contrasting patterns of polymorphism and divergence between these categories of mutations. I use available DNA sequences to identify the action of natural selection and to estimate the intensity of selection at silent sites in the *D. melanogaster* subgroup.

METHODS AND RESULTS

Identifying preferred codons in *D. melanogaster*: Natural selection for a preferred codon will increase its frequency relative to other members of a synonymous family (Figure 1). The "scaled chi square" of SHIELDS *et al.* (1988) provides a measure of the degree of codon usage bias for a given gene. Values of chi square for deviations from an A/T content of 60%, the average base content of *D. melanogaster* introns (SHIELDS *et al.* 1988; MORIYAMA and HARTL 1993), are calculated for each synonymous family, excluding the codon family under analysis. The sum of these chi-square values is divided by the total number of codons in a gene to give a measure of codon bias independent of gene length, referred to as scaled chi square (at60). Preferred codons are those that increase in frequency as a function of codon bias (Table 1). The small number of available protein-coding sequences and lack of silent DNA divergence between *D. melanogaster* and *D. simulans* (K_s averages ~10%) preclude this method from detecting differences in codon preference between these sibling species. In the following analyses, codon preference will be assumed to be conserved between *D. melanogaster* and *simulans*.

Polymorphism and divergence of synonymous DNA changes: Comparison of ratios of polymorphism to divergence between classes of closely linked sites can identify differences in the fitness effects of mutations in DNA (MCDONALD and KREITMAN 1991; SAWYER and HARTL 1992). Ratios of polymorphism and divergence, referred to as r_{pd}, are compared for synonymous mutations at the *Adh*, *Adhr*, *Amy*, *per* and *Pgi* loci in *D. melanogaster* and *simulans* (see Table 2 for references). Two or more alleles of these genes have been sequenced in both species, and each has been sequenced in at least one other species within the *D. melanogaster* subgroup. Outgroup sequences allow inference of the direction of synonymous DNA changes within or between *D. melanogaster* and *simulans* (Figure 2). Table 2 shows 2 × 2 contingency tables comparing numbers of polymorphic and fixed mutations for the two classes of synonymous changes. Selection for codon bias predicts a higher ratio of polymorphism to divergence for deleterious unpreferred mutations than for advantageous preferred silent changes. Homogeneity in these tables indicates no difference in the fitness effects of the classes of mutations (neutrality of both classes is a subset of this null hypothesis). Because there is no prediction for the fitness effect of mutations occurring within classes, preferred to preferred and unpreferred to unpreferred changes are not included in the analysis. To determine the statistical significance of the deviations in these contingency tables, data for the five loci were combined using the MANTEL–HAENSZEL test with a correction for continuity (MANTEL and HAENSZEL 1959; MANTEL 1963). This procedure takes into account both the magnitude and direction of deviations in testing the null hypothesis of no difference in r_{pd}'s between the two categories of mutations.

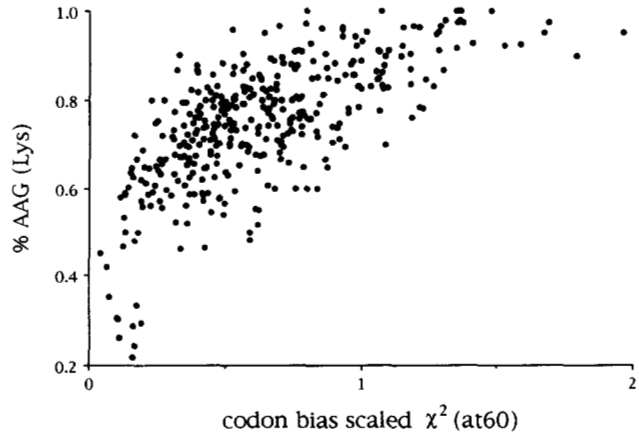

FIGURE 1.—Biased codon usage for lysine codons in *D. melanogaster*. Frequencies of AAG within lysine codons (AAA and AAG) are plotted against scaled chi square (at60) values for genes with a minimum of 20 lysine codons among 575 *D. melanogaster* genes drawn from GenBank. The increase in the frequency of AAG among lysine codons as a function of the intensity of selection for codon usage bias identifies it as a preferred codon. In this analysis, synonymous changes from AAA → AAG are classified as preferred changes and mutations in the opposite direction, AAG → AAA, are considered unpreferred.

The ratio of the number of synonymous changes segregating within *D. simulans* to the number of mutations that have fixed in this lineage is significantly higher for unpreferred than for preferred synonymous mutations (MANTEL-HAENSZEL test, $z = 3.00$, $P = 0.001$). r_{pd}s of preferred and unpreferred synonymous changes differ in the direction predicted by the selection-mutation-drift model in *D. simulans*. r_{pd}s for the two classes of synonymous mutations do not show a significant difference in *D. melanogaster* ($z = 0.44$, $P = 0.33$); the null hypothesis of no fitness differences between classes of synonymous changes cannot be rejected.

Synonymous fixations in *D. melanogaster* and *D. simulans*: If the frequency of preferred codons in a given gene is maintained under relatively constant selection intensity, mutation rates and population sizes, the number of unpreferred fixations will equal that of preferred fixations. Departures from this expectation show that codon bias is either increasing or

TABLE 1

Preferred codons in *D. melanogaster*

Amino acid	Codon	χ^2	Amino acid	Codon	χ^2	Amino acid	Codon	χ^2	Amino acid	Codon	χ^2
Phe	TTT		Ser	TCT		Tyr	TAT		Cys	TGT	
	TTC*	547.5		TCC*	169.7		TAC*	314.7		TGC*	142.0
Leu	TTA			TCA		Ter	TAA		Ter	TGA	
	TTG			TCG*	42.7		TAG		Trp	TGG	
Leu	CTT		Pro	CCT		His	CAT		Arg	CGT	7.0
	CTC*	47.7		CCC*	317.1		CAC*	157.2		CGC*	225.0
	CTA			CCA		Gln	CAA			CGA	
	CTG*	496.7		CCG	8.7		CAG*	517.3		CGG	
Ile	ATT		Thr	ACT		Asn	AAT		Ser	AGT	
	ATC*	627.4		ACC*	398.6		AAC*	339.4		AGC*	184.9
	ATA			ACA		Lys	AAA		Arg	AGA	
Met	ATG			ACG	0.7		AAG*	542.2		AGG	
Val	GTT		Ala	GCT		Asp	GAT		Gly	GGT	
	GTC*	134.7		GCC*	568.4		GAC*	185.5		GGC*	166.3
	GTA			GCA		Glu	GAA			GGA	
	GTG*	67.4		GCG			GAG*	593.4		GGG	

Frequencies of each codon within its synonymous family were compared to the scaled χ^2 (at60) values (excluding the synonymous family being analyzed) of 575 *D. melanogaster* genes. Serine codons are divided into a fourfold and twofold degenerate family. χ^2 values (1 d.f.) of logistic regressions for codons whose frequencies show a positive slope as a function of scaled χ^2 (at60) are shown. Twenty-two codons showing a significant relationship ($P < 0.05$) using the sequential Bonferroni test (RICE 1989) are defined as preferred codons and are marked with *. Data available from author.

decreasing. Table 3 shows the pooled numbers of unpreferred and preferred synonymous fixations compared with the expectation of equal numbers in the two classes for the five loci analyzed above. Because comparison of synonymous *fixations* does not require controlling for differences in the recent evolutionary histories of different regions of the genome, data can be pooled across loci. The two classes of synonymous changes have fixed at approximately equal rates in *D. simulans* ($G = 0.32$, $P > 0.5$); the null hypothesis of constant selection intensity for codon bias cannot be rejected. In *D. melanogaster*, unpreferred fixations outnumber preferred fixations by over sixfold ($G = 19.3$, $P = 10^{-5}$), supporting a relaxation of selection at silent sites for the five genes analyzed.

Rates of unpreferred and preferred fixations can also be contrasted between lineages without assigning the direction of mutations using outgroup sequences. If rates of unpreferred and preferred synonymous substitution are the same in the two lineages, then the number of positions where the *D. melanogaster* gene encodes a preferred codon and *D. simulans* encodes an unpreferred codon (mp/su) will not differ from the number of positions where codon usage differences are reversed between species (mu/sp). Table 4 shows the number of mp/su and mu/sp codons at homologous codons where amino acids are conserved between species. WILCOXON's (1945) signed rank test was used to determine whether codon usage between these species departs from a null hypothesis of no difference in the two categories (mu/sp = mp/su). The 28 genes show a significant excess of codons encoding an unpreferred codon in *D. melanogaster* and a preferred codon in *D. simulans* (mu/sp) (WILCOXON's signed rank test, $P = 10^{-4}$). Relaxation of selection at silent sites in *D. melanogaster* is not restricted to the five genes previously considered but appears to be genome-wide.

Estimating selection coefficients at silent sites: Selection coefficients on silent DNA changes can be estimated from observed levels of polymorphism and divergence using the method of SAWYER and HARTL (1992). r_{pd} for nonneutral changes is a function of three parameters: divergence time scaled to N_e generations, t_{div}, the number of genes sampled from a population, m, and the product of effective population size and selection coefficient, $N_e s$. The mutation rate, μ, does not affect the ratio of polymorphism to divergence. t_{div} can be estimated from r_{pd} for neutral mutations and m using the sampling formulae of SAWYER and HARTL (1992, see also HUDSON 1990):

$$t_{div} = \frac{L(m)}{r_{pd}} - \frac{1}{m},$$

where

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

t_{div} is calculated from putatively neutral intron polymorphism and divergence using alleles of *per* and *Pgi* in *D. simulans* shown in Table 2. These genes were chosen because they show similar levels of codon bias and because they provide the largest sample sizes from which to estimate selection intensity. Larger sample sizes allow observed r_{pd} values for unpreferred changes to be used with greater confidence than r_{pd} for rarer preferred changes. Results are given for the two genes independently and for the pooled data. Because the analysis requires the same number of alleles sequenced for each gene, a single allele (L27547) was eliminated at random for *Pgi* so that $m = 6$ for the pooled data. Divergence in the *D. simulans* lineage is calculated as half the number of fixed differences between *D. melanogaster* and *D. simulans*. Putatively neutral intron r_{pd} values of 5/5.5 and 17/14 give t_{div} values of 2.3 and 1.9 at *per* and *Pgi*, respectively. The combined (minus one allele for *Pgi*) r_{pd} of 22/19.5 gives t_{div} of 1.9.

These values are used to calculate $\gamma = N_e s$ for unpreferred

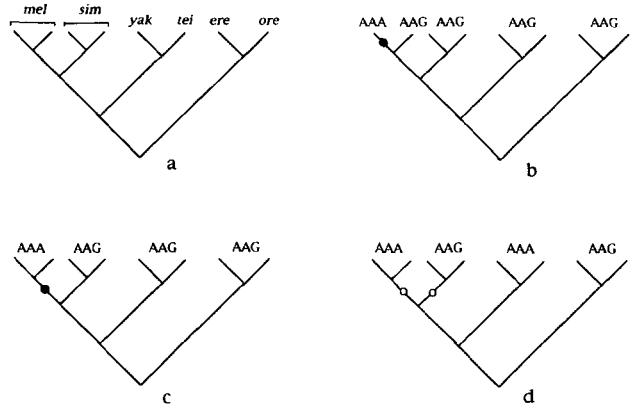

FIGURE 2.—Inferring the direction of synonymous changes. Tree a shows phylogenetic relationships within the *D. melanogaster* subgroup proposed by Lachaise *et al.* (1988) and Jeffs *et al.* (1994). Evolutionary trees connect a single codon that has changed within or between two alleles each of *D. melanogaster* and *D. simulans* and one allele each of *Drosophila yakuba*, *Drosophila teisseri*, *Drosophila erecta* and *Drosophila orena* labeled *mel*, *sim*, *yak*, *tei*, *ere* and *ore*, respectively. The most parsimonious change for the codons of tree b is an unpreferred polymorphism, AAG → AAA, in the *D. melanogaster* lineage. The location of the mutation on the tree is marked with a dot. The most parsimonious change in tree c is an unpreferred fixation, AAG → AAA, in *D. melanogaster*. Tree d shows a case for which the direction of a synonymous mutation is ambiguous (multiple trees give the least number of changes). A total of 28 such cases are not included in the analysis. A single case in which two silent mutations were found segregating at a single site was not included in the analysis because the direction of the mutations could not be inferred. Note that the relative positions of the *ere*/*ore* and *tei*/*yak* lineages do not affect inference of the direction of synonymous mutations.

synonymous DNA changes from the SAWYER-HARTL (1992) formulae:

$$r_{pd} = \frac{F(m)}{t_{div} + G(m)},$$

where

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

and

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

Numerical solutions to these equations for the t_{div} values given above and r_{pd}s of 25/3 and 13/1 (Table 2) give $N_e s$ of −2.3 and −2.4 for unpreferred changes at *per* and *Pgi*, respectively. The combined r_{pd} of 37/4 (elimination of one allele of *Pgi* reduces the number of polymorphic sites by one) gives $N_e s$ of −2.2.

These equations were also used to obtain a bootstrap maximum likelihood confidence interval for the estimate of $N_e s$ for the pooled data. For each iteration, the four variables, numbers of polymorphic and fixed mutations for intron and unpreferred silent changes, were resampled from Poisson distributions with the observed values as means. $N_e s$ was reestimated numerically from the resampled data in each iteration.

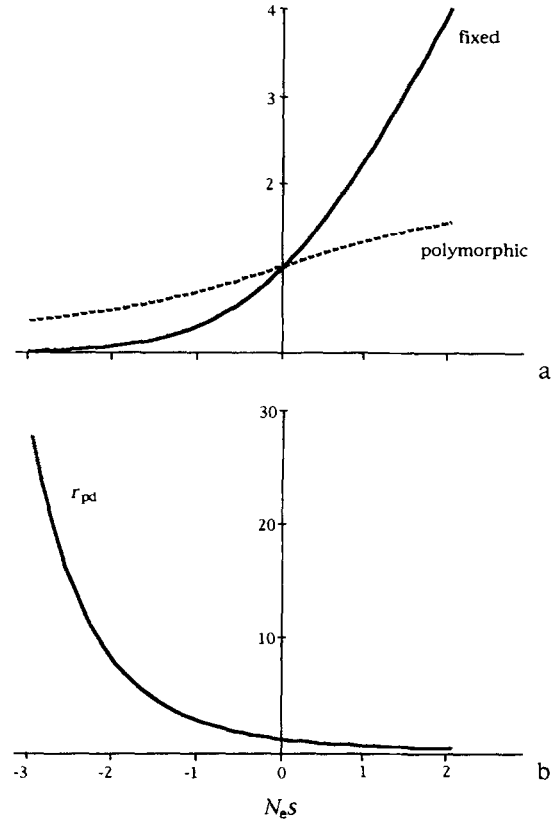

FIGURE 3.—Polymorphism and divergence of nearly neutral mutations. (a) Expected numbers of polymorphic (– – –) and fixed (——) mutations in a sample of DNA sequences as a function of $N_e s$, the product of effective population size and selection coefficient (see KIMURA 1983). SAWYER and HARTL's sampling formulae (1992) were used to calculate expected numbers of polymorphic and fixed changes relative to the expected numbers for neutral changes. Parameter values (t_{div} = 1.9, m = 6, see text) were calculated for the *per* and *Pgi* genes in *D. simulans*. (b) The ratio of expected numbers of polymorphic and fixed differences, r_{pd}, as a function of $N_e s$ for the same parameter values.

The number of sampled sites for the introns was 833, the total number of aligned intron sites examined (excluding, for each intron, the two invariable 5′ and 3′ bases). For the coding sequence, the number of sampled "unpreferred sites," 488, was calculated as the number of single base mutations at preferred codons that give rise to unpreferred codons divided by three (because only one mutation can be observed at a given base). In 10,000 iterations, the median value was $N_e s = -2.3$. The mid-95% interval ($-3.6 < N_e s < -1.3$) was taken as the confidence interval for the estimate from the data.

DISCUSSION

Inferring natural selection from ratios of polymorphism to divergence: The effectiveness of natural selection in determining the fate of a mutation depends on the product of effective population size and selection coefficient, $N_e s$ (KIMURA 1983). KIMURA's studies of the behavior of mutations in the neighborhood of neutral-

TABLE 2

Polymorphism and divergence of synonymous DNA changes in *D. melanogaster* and *D. simulans*

Gene		*mel* No. of alleles	*mel* Unpref	*mel* Pref	*sim* No. of alleles	*sim* Unpref	*sim* Pref
Adh	Poly	12	11	1	6	5	1
	Fixed		1	0		0	0
Adhr	Poly	12	0	0	5	9	3
	Fixed		5	1		1	1
Amy	Poly	18	32	1	2	18	0
	Fixed		5	1		0	0
per	Poly	6	13	0	6	25	0
	Fixed		9	0		3	2
Pgi	Poly	8	1	0	7	13	4
	Fixed		12	3		1	4
MANTEL-HAENSZEL test (one tailed)			$z = 0.44$ $P = 0.33$			$z = 3.00$ $P = 0.001$	

2×2 contingency tables are shown for unpreferred and preferred synonymous mutations found polymorphic and fixed in *D. melanogaster* and *simulans* at the *Adh, Adhr, Amy, per* and *Pgi* loci. The numbers of alleles examined in each species are given in parentheses. GenBank/EMBL DNA sequence library (release 81.0) accession numbers or references are as follows: *Adh* (*mel:* M17827, M17828, M19547, M17830-M17837, M22210, *sim:* M19276, M19263, X57361-X57364, *ere:* X54116, *ore:* M37837, *yak:* X57365-X57376, *tei:* X54118). *Adhr* (*mel:* (KREITMAN and HUDSON 1991), *sim:* M36581 (SUMNER 1991), *ere:* X54116, *yak:* X54120, *tei:* X54118). *Amy* (*mel:* L22716-L22735, *sim:* D17733, D17734, *ere:* D17727, D17728, *ore:* D21128, D21129, *yak:* D17737, D17738, *tei:* D17735, D17736). *Per* (*mel:* L07817-L07819, L07821, L07823, L07825, *sim:* L07826-L07832), *yak:* X61127). *Pgi* (*mel:* L27539-L27546, *sim:* L27547-L27552, *yak:* L27673-L27685, *tei:* J. H. MCDONALD, personal communication). *mel, sim, yak, tei, ere* and *ore* refer to *D. melanogaster, D. simulans, D. yakuba, D. teisseri, D. erecta* and *D. orena,* respectively.

ity ($|N_e s| \sim 1$) show that the probability that a nonneutral change will rise to an intermediate frequency within a population is less affected by selection than the probability that it will go to fixation (KIMURA 1983, Figure 3a). The *ratio* of polymorphism within a species to divergence between species changes dramatically as a function of $N_e s$ and is sensitive to even very weak selection (Figure 3b).

TABLE 3

Synonymous DNA fixations in *D. melanogaster* and *D. simulans*

	mel Unpref	*mel* Pref	*sim* Unpref	*sim* Pref
Observed	32	6	5	7
Expected	19	19	6	6
G-test	$G = 19.3$ $P = 10^{-5}$		$G = 0.32$ $P > 0.5$	

2×2 contingency tables are shown for unpreferred and preferred synonymous fixations in *D. melanogaster* and *simulans* at the *Adh, Adhr, Amy, per* and *Pgi* loci (data from Table 2 are pooled for each species). A *G*-test for goodness of fit with the Williams correction for continuity (SOKAL and ROHLF 1981) was used to test for departures from the expectation of equal numbers in the two classes.

MCDONALD and KREITMAN (1991) first suggested the comparison of ratios of polymorphism to divergence to identify differences in $N_e s$ between classes of closely linked mutations in DNA. Selection-mutation-drift at silent sites in Drosophila predicts the action of purifying selection on unpreferred synonymous mutations ($N_e s < 0$) and directional selection favoring preferred changes ($N_e s > 0$). Higher ratios of polymorphism to divergence for unpreferred changes than for preferred changes confirms this model of codon selection in the *D. simulans* lineage (Table 2).

The comparison of ratios of polymorphism to divergence is relatively free of assumptions (MCDONALD and KREITMAN 1991). Because r_{pd} values are compared between two classes of closely linked mutations, the validity of the statistical test does not require that populations have reached equilibrium or that all sites are evolving independently. If the two types of mutations are interspersed within genes, different evolutionary histories of local regions should have an equivalent effect on the number of polymorphic sites and, consequently, the r_{pd}'s of both classes.

Mutation rates are not assumed to be equal between the categories of changes but are assumed to be constant within each class within each lineage. KLIMAN and HEY (1994) recently found a correlation ($r^2 = 0.101$) between putatively neutral intron base composition and

TABLE 4

Synonymous codon usage differences between *D. melanogaster* and *D. simulans*

Gene	Codon bias scaled χ^2 (at60)	Codon differences		
		mp/su	mu/sp	mu/sp − mp/su
ac	0.29	4	9	+5
Acp26Aa	0.20	4	5	+1
Acp26Ab	0.54	0	2	+2
Act88F	1.13	4	13	+9
Adh	1.09	1	8	+7
Adhr	0.22	4	11	+7
Amy	1.20	12	14	+2
ase	0.17	4	6	+2
bcd	1.19	0	0	
ci	0.34	6	5	−1
cta	0.21	6	4	−2
Est-6	0.20	10	25	+15
g6pd	1.03	6	19	+13
GstD1	1.45	1	8	+7
Hsc70-1	0.61	5	8	+3
Hsp83	0.89	4	9	+5
Mlc1	0.83	0	3	+3
MtnA	1.14	0	0	
per	0.74	16	19	+3
Pgd	0.73	21	19	−2
Pgi	0.75	11	19	+8
ref(2)P	0.25	12	10	−2
sala	0.58	5	6	+1
Sod	0.74	3	8	+5
su(f)	0.20	13	25	+12
tra	0.28	6	6	
Yp2	0.96	5	8	+3
z	0.52	4	9	+5
WILCOXON's signed rank test			$z = 3.81$ $P = 10^{-4}$	

Comparison of codons conserved for amino acids but differing at silent positions between *D. melanogaster* and *simulans* in 28 genes. Gene names follow FlyBase (1993). Codon bias scaled χ^2 (at60), averaged between species, are given for each gene. The number of condons where the *D. melanogaster* gene encodes the preferred codon and *D. simulans* encodes an unpreferred codon (mp/su) is compared with the number of positions where the synonymous differences are reversed between species (mu/sp). For loci with multiple alleles sequenced within a species a single sequence was chosen at random for this analysis. GenBank/EMBL DNA sequence library (release 81.0) accession numbers or references are as follows: *ac* (*m:* M17120, *s:* M17120), *Acp26Aa* (*m:* X70888, *s:* X70888), *Acp26Ab* (*m:* X70888, *s:* X00607), *Act88F* (*m:* M18830, *s:* M87274), *Adh* (*m:* M17827, *s:* M19276), *Adhr* (*m:* KREITMAN and HUDSON 1991), *s:* M17120), *Amy* (*m:* L22724, *s*-D17733), *ase* (*m:* X52892, *s:* J. HEY, personal communication), *bcd* (*m:* X07870, *s:* M32123), *ci* (*m:* X54360, *s:* (BERRY *et al.* 1991)), *cta* (*m:* M94285, *s:* M. WAYNE, personal communication), *Est-6* (*m:* J04167, *s:* L10670), *G6pd* (*m:* L13880, *s:* L13876), *GstD1* (*m:* X14233, *s:* M84577), *Hsc70-1* (*m:* L01501, *s:* J01089), *Hsp83* (*m:* X03810, *s:* X03811), *Mlc1* (*m:* K01567, *s:* L08051), *MtnA* (*m:* M12964, *s:* M55407), *per* (*m:* L07817, *s:* L07826), *Pgd* (*m:* M80598, *s:* U02288), *Pgi* (*m:* L27539, *s:* L27547), *ref(2)P* (*m:* X16993, *s:* M. WAYNE, personal communication), *sala* (*m:* X57474, *s:* M21227), *Sod* (*m:* M24421, *s:* X15685), *su(f)* (*m:* X62679, *s:* L09193), *tra* (*m:* M17478, *s:* X66930), *Yp2* (*m:* L14421, *s:* L14426), *z* (*m:* L13043, *s:* L13049). *m* and *s* refer to *D. melanogaster* and *D. simulans*, respectively.

the base content of silent sites in *D. melanogaster*. Their finding suggests that variation in mutation rates and/or biases contributes to some of the variation in codon usage found between genes. Such biases, if they have remained constant since the split between *D. melanogaster* and *D. simulans*, will have an equal impact on the number of sampled polymorphic and fixed differences within each class of changes and should not affect comparison of their ratios. The relative contributions of selection and mutational bias in governing the base composition of Drosophila genes, however, is not addressed by this analysis.

Because this analysis requires assignment of the direction of mutations, parallel changes or errors in determining preferred codons can result in misclassification of synonymous changes. Such errors, however, will not bias the statistical test if the evolutionary dynamics of the two classes of mutations do not differ. Significantly higher ratios of polymorphism to divergence for unpreferred than for preferred synonymous mutations

in *D. simulans* are difficult to explain in the absence of natural selection.

Weak selection at silent sites: The lack of a significant difference in r_{pd} between preferred and unpreferred synonymous changes suggests that selection at silent sites may be less effective in *D. melanogaster*. Comparison of synonymous fixation rates reveals an excess of unpreferred fixations in *D. melanogaster*, confirming a genome-wide relaxation of selection at silent sites in this lineage (Tables 3 and 4). Natural selection may play a role in silent DNA evolution in *D. melanogaster*, but codon bias has decreased in this lineage since its split from the common ancestor to *D. simulans*.

Higher mutation rates, smaller selection coefficients or a smaller effective population size could explain the reduced efficacy of selection at silent sites in *D. melanogaster*. Estimates of DNA heterozygosities are approximately three- to sixfold lower in *D. melanogaster* than in *D. simulans*, suggesting a difference in effective population size of the same magnitude (AQUADRO 1992). Such a difference in N_e could explain differences in codon usage between *D. melanogaster* and *simulans* if $N_e s$ is close to unity for synonymous mutations. The sampling equations of SAWYER and HARTL (1992) allow $N_e s$ to be estimated from observed ratios of polymorphism to divergence under assumptions of haploid populations at equilibrium and independent evolution at all sites. The sensitivity of r_{pd} to $N_e s$ illustrates the power of this method in estimating selection coefficients for slightly deleterious mutations (Figure 3b).

The SAWYER-HARTL formulae give an estimate of $N_e s = -2.2$ for the average selection intensity acting on unpreferred synonymous mutations at the *per* and *Pgi* loci in the *D. simulans* lineage. A maximum likelihood approach gives a relatively small confidence interval around this estimate ($-3.6 < N_e s < -1.3$). Although violation of the equilibrium assumption may have a substantial effect on these estimates, it is encouraging that independent replicates of codon bias evolution under similar levels of selection (codon bias scaled chi square (at60) is 0.74 and 0.75 for *per* and *Pgi*, respectively) yield estimates of $N_e s$ of -2.3 at *per* and -2.4 at *Pgi*.

These estimates of $|N_e s| \sim 2$ for synonymous DNA changes in *D. simulans* can explain the reduced efficacy of selection ($|N_e s| < 1$) in the three- to sixfold smaller N_e of *D. melanogaster*. However, because current DNA heterozygosities reflect population history for up to about four N_e generations, it is possible that differences in effective population sizes between these species have been much greater in the past. In addition, higher mutation rates or smaller selection coefficients cannot be excluded as factors contributing to relaxed selection for codon bias. Genome-wide differences in codon usage bias between species could also arise from differences in selection intensity for metabolic efficiency or growth rates. In the absence of evidence for such a difference between *D. melanogaster* and *simulans*, a smaller effective population size appears to be sufficient to account for the relaxation of selection at silent sites in *D. melanogaster*.

Selection coefficients for synonymous mutations are on the order of $1/N_e$ in Drosophila. OHTA's (1973, 1992) studies of nearly neutral mutations reveal two important properties of weakly selected mutations ($|N_e s| \sim 1$) that differentiate their evolutionary dynamics from that of strictly neutral changes ($N_e s = 0$). First, because the relative contributions of stochastic (genetic drift) and deterministic (natural selection) forces depend critically on species effective population size, substitution rates of weakly selected mutations are expected to vary between even closely related species (OHTA 1992). Differences in silent DNA evolution between the morphologically and ecologically almost indistinguishable sibling species *D. melanogaster* and *simulans* appear to confirm the sensitivity of synonymous DNA evolution to N_e. OHTA (1973) and OHTA and TACHIDA (1990) have also noted that, because slightly deleterious mutations occasionally fix by chance even in larger populations, the maintenance of a given level of fitness requires a continuous input of adaptive substitutions. In the absence of positive selection for preferred changes at silent sites, a gradual accumulation of unpreferred fixations will drive synonymous codon usage toward mutational equilibrium. The maintenance of major codon preference for nearly neutral synonymous DNA changes requires a slow, but constant, rate of "compensatory" adaptive fixations.

Constant and context-dependent selection models of codon bias: The estimates of $N_e s$ presented here reflect an average for the class of mutations under investigation and do not address the issue of whether selection intensity varies within or between genes. The simplest model of silent DNA evolution assumes constant selection intensity for preferred codons independent of the frequency of preferred codons in a given gene. However, constant selection models require a very restricted range of selection intensity (about an order of magnitude) to explain the range of codon bias observed in the genes of species exhibiting major codon preference (EYRE-WALKER 1994). If $N_e s$ falls much below unity, genetic drift will overwhelm deterministic forces and codon usage will show little bias. If $N_e s$ is greater than three or four (depending on mutational biases), all codons will be fixed for preferred codons. HARTL *et al.*'s (1994) estimate of $N_e s = -1.3$ in the *E. coli gnd* gene based on the frequency distribution of silent DNA polymorphism is indistinguishable from that estimated here in Drosophila. However, the effective population sizes of these species are thought to vary by 20-fold (AYALA and HARTL 1993; HARTL *et al.* 1994). Under the constant selection model, selection coefficients at silent sites must fall precisely within an order of magnitude of

$1/N_e$ in both species to produce the range of observed codon bias.

EYRE-WALKER (1994) has developed a test of the constant selection model for codon bias. Constant selection intensity predicts that the proportion of saturation at silent sites (a measure of divergence) in a given gene should be independent of the degree of codon bias. Both *E. coli* and yeast show a strong negative correlation between the proportion of saturation and the frequency of preferred codons; the simple constant selection model of codon bias evolution can be rejected. EYRE-WALKER (1994) suggests that constraints of ribosome binding and mRNA secondary structure (both of which are supported by experimental evidence) must be invoked to explain codon bias evolution. Although this may be valid in some species, a substantial contribution of such conflicting selection pressures predicts that the average fitness effects of unpreferred and preferred synonymous changes should not differ as observed in Drosophila.

KIMURA's (1981) stabilizing selection model of codon bias evolution is also inconsistent with findings presented here. Under this model, the optimal fitness state occurs when the relative frequencies of synonymous codons match those of the cognate tRNAs. This scenario differs from the selection-mutation-drift model because the fitness effect of a silent mutation will depend on the current state of codon bias in the genome with respect to the relative frequencies of available tRNAs. Unpreferred (and preferred) mutations may be either positively or negatively selected depending on whether they shift codon usage toward or away from the optimal fitness state. KIMURA's theory does not predict consistent fitness difference between unpreferred and preferred synonymous mutations.

A context-dependent selection model of codon evolution may be necessary to account for both fitness diferences between classes of silent changes and similar patterns of codon bias in species with different population sizes. The fitness benefit of encoding preferred codons may decrease as the number of preferred codons increases in a given gene [see LI's (1987) discussion of "synergistic" codon selection]. Under this model, the number of preferred codons reaches an equilibrium between mutation pressure and the intensity of selection (given a certain frequency of preferred codons). At equilibrium, weak purifying selection against unpreferred changes and directional selection favoring preferred substitutions will maintain a given frequency of preferred codons. Differences in codon bias between species depend on effective population sizes, N_e, and on how dramatically selection coefficients, s, decrease as a function of the frequency of preferred codons. If the fitness benefit to encoding preferred codons decreases sharply as a function of codon bias, then species with vastly different effective population sizes may show relatively small differences in codon bias.

Both constant and context-dependent selection models can account for observed levels of codon bias with weak selection intensity (LI 1987). Because codon bias in *D. simulans* appears to have reached equilibrium (Table 3), the analyses presented here do not distinguish between these models. One experimental observation, however, suggests that the fitness effect of codon usage is not independent of other codons in the same gene. VARENNE *et al.* (1989) found that unpreferred codons clustered together cause a greater decrease in translational elongation rate than a larger number of unpreferred codons spaced more widely apart in the *E. coli cat* gene. The context of unfavorable codons may play an important role in determining their effect on translation and, consequently, their effect on fitness.

Synonymous DNA changes and tests of adaptive protein evolution: Weak selection on synonymous DNA changes has important implications for the use of silent mutations in tests of evolutionary hypotheses. MCDONALD and KREITMAN (1991), in their original formulation of the comparison of ratios of polymorphism to divergence, compared r_{pd} between silent and replacement changes. Significantly lower ratios of polymorphism to divergence for replacement changes at the *Adh* (MCDONALD and KREITMAN 1991), *G6pd* (EANES *et al.* 1993) and *jgw* (LONG and LANGLEY 1993) genes in the *Drosophila melanogaster* subgroup could be interpreted as evidence of adaptive protein evolution if synonymous changes are neutral. However, in genes showing biased codon usage, most codons are ancestrally in the favored state. Because the majority of newly arising mutations are from preferred to unpreferred codons in such genes, r_{pd} for the pooled class of silent changes will reflect purifying selection on the larger number of unpreferred changes ($N_e s < 0$). A significantly lower r_{pd} at replacement sites implies less negative selection coefficients than at silent sites but does not allow inference of whether selection coefficients are negative, positive or zero. The sign of $N_e s$ can be determined if replacement changes are compared separately with the two classes of silent mutations. If amino acid changes show a lower ratio of polymorphism to divergence than preferred silent changes, then positive selection coefficients can be inferred. If r_{pd} for replacement changes is significantly higher than for unpreferred silent changes, then $N_e s$ must be negative. The power to detect purifying or directional selection in protein evolution will be greatest in genes showing low codon bias, where sample sizes will be greater for both classes of synonymous change and $N_e s$ for unpreferred and preferred mutations will converge toward zero.

Selection on synonymous DNA changes must also be taken into account when comparing levels of polymorphism and divergence using the approach of HUDSON

et al. (1987). The HKA method tests the null hypothesis of equal ratios of polymorphism and divergence between different regions of the genome. An excess of polymorphism in a given region has be interpreted as evidence for balancing selection (HUDSON *et al.* 1987), whereas too little polymorphism has been argued as evidence for genetic "hitchhiking" associated with a recent adaptive fixation (reviewed in AQUADRO and BEGUN 1993). Critical to the validity of these interpretations is the assumption of neutrality of the mutations compared between regions; the ratio of polymorphism to divergence remains constant for neutral mutations undergoing similar evolutionary histories. As KIMURA (1983) has shown, and I confirm for synonymous changes, weak selection can substantially alter ratios of polymorphism to divergence from that expected under neutrality (see also CHARLESWORTH 1994). A highly codon-biased gene will show higher levels of polymorphism to divergence at silent sites than a low-biased gene or neutral region where $N_e s$ is smaller.

Molecular population genetics and synonymous DNA changes: Although studies attempting to infer evolutionary processes from DNA sequence data have concentrated on protein evolution, silent changes offer some important advantages for this form of statistical inference. Patterns of codon usage suggest subdivision of synonymous mutations into deleterious and advantageous classes, and the action of natural selection can be identified from the comparison of categories of silent changes. Replacement mutations, however, are generally treated as a single category in statistical analyses (although see HUGHES *et al.* 1990). The relative contributions of purifying and directional selection may be masked by pooling together advantageous, neutral and deleterious amino acid mutations. In the absence of a methodology to predict the fitness effect of particular amino acid changes, our understanding of protein evolution may not compare with the model presented here for silent sites in DNA.

I am grateful to MARTY KREITMAN for valuable discussions and guidance throughout this work. JODY HEY, JOHN H. MCDONALD and MARTA WAYNE kindly provided unpublished DNA sequences. I also thank BRIAN CHARLESWORTH, JAMES CROW, JODY HEY, JOHN H. MCDONALD, THOMAS NAGYLAKI, STEVEN ORZACK, ELI STAHL, STANLEY SAWYER, CHUNG-I WU and two anonymous reviewers for suggestions and discussion. Differences in codon usage bias between *D. melanogaster* and *simulans* have been noted independently by JODY HEY (personal communication). H.A. is a Howard Hughes Medical Institute Predoctoral Fellow. This research was also supported by National Institutes of Health grant GM-39355 to MARTIN KREITMAN.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1992 Polymorphism and divergence in the *Mst26A* male accessory gland gene region in Drosophila. Genetics **132:** 755–770.

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

AQUADRO, C. F. 1992 Why is the genome variable? Insights from Drosophila. Trends Genet. **8:** 355–362.

AQUADRO, C. F., and D. J. BEGUN, 1993 Evidence for and implications of genetic hitchhiking in the Drosophila genome, pp. 159–178 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, MA.

AYALA, F. J., and D. L. HARTL, 1993 Molecular drift of the *bride of sevenless* (*boss*) gene in *Drosophila*. Mol. Biol. Evol. **10:** 1030–1040.

BENNETZEN, J. L., and B. D. HALL, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

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

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

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.

CHARLESWORTH, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–228.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

EYRE-WALKER, A., 1994 The evolution of synonymous codon use—testing the mutation-selection-drift hypothesis. Mol. Biol. Evol. (in press).

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

HARTL, D. L., E. N. MORIYAMA and S. SAWYER, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Series in Ecology and Evolution*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUGHES, A. L., T. OTA and M. NEI, 1990 Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7:** 515–524.

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* translation system. J. Mol. Biol. **151:** 389–409.

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

JEFFS, P. S., E. C. HOLMES, and M. ASHBURNER, 1994 The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **11:** 287–304.

KIMURA, M., 1981 Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc. Natl. Acad. Sci. USA **78:** 5773–5777.

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

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

KLIMAN, R. M., and J. HEY, 1993b DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

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.

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

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogater* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LACHAISE, D., M. L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

LI, W.-H., 1987 Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

LONG, M., and C. H. LANGLEY, 1993 Natural selection and the origin of *jingwei*, a processed functional gene in *Drosophila*. Science **260:** 91–95.

MANTEL, N., 1963 Chi-squared tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. J. Am. Stat. Assoc. **58:** 690–700.

MANTEL, N., and W. HAENSZEL, 1959 Statistical aspects of the analysis of data from the retrospective analysis of disease. J. Natl. Cancer Inst. **22:** 719.

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

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

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

OHTA, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Sys. **23:** 263–286.

OHTA, T., and H. TACHIDA, 1990 Theoretical Study of Near Neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219–229.

RICE, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

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

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

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, 1981 *Biometry*. W. H. Freeman and Co., New York.

SUMNER, C., 1991 Nucleotide polymorphism in the *Alcohol dehydrogenase Duplicate* locus of *Drosophila simulans:* implications for the neutral theory. Undergraduate thesis. Princeton University, Princeton, NJ.

VARENNE, S., C. BATY, H. VERHEIJ, D. SHIRE and C. LAZDUNSKI, 1989 The maximum rate of gene expression is dependent on the downstream context of unfavorable codons. Biochimie **71:** 1221–1229.

WILCOXON, F., 1945 Individual comparisons by ranking methods. Biometrics Bull. **1:** 80–83.

Communicating editor: A. G. CLARK