Natural Selection and the Frequency Distributions of "Silent" DNA Polymorphism in Drosophila

Hiroshi Akashi* and Stephen W. Schaeffer†

**Section of Evolution and Ecology, University of California at Davis, Davis, California 95616, and †Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pennsylvania 16802*

Manuscript received May 11, 1996
Accepted for publication February 10, 1997

ABSTRACT

In *Escherichia coli, Saccharomyces cerevisiae,* and *Drosophila melanogaster,* codon bias may be maintained by a balance among mutation pressure, genetic drift, and natural selection favoring translationally superior codons. Under such an evolutionary model, silent mutations fall into two fitness categories: preferred mutations that increase codon bias and unpreferred changes in the opposite direction. This prediction can be tested by comparing the frequency spectra of synonymous changes segregating within populations; natural selection will elevate the frequencies of advantageous mutations relative to that of deleterious changes. The frequency distributions of preferred and unpreferred mutations differ in the predicted direction among 99 alleles of two *D. pseudoobscura* genes and five alleles of eight *D. simulans* genes. This result confirms the existence of fitness classes of silent mutations. Maximum likelihood estimates suggest that selection intensity at silent sites is, on average, very weak in both *D. pseudoobscura* and *D. simulans* ($|N_e s| \approx 1$). Inference of evolutionary processes from within-species sequence variation is often hindered by the assumption of a stationary frequency distribution. This assumption can be avoided when identifying the action of selection and tested when estimating selection intensity.

UNDER KIMURA and OHTA's view of molecular evolution, gene frequency changes within species and the accumulation of fixed differences between species reflect a single process of mutation, genetic drift, and (for some mutations) unidirectional selection (KIMURA and OHTA 1971; KIMURA 1983). This theoretical framework has been extremely useful for developing statistical methods both to establish the role of natural selection in shaping genetic variation and to estimate the magnitude of parameters governing molecular evolution.

If polymorphism is a phase of molecular divergence, then the frequency distribution of mutations segregating within a species constitutes a "snapshot" of the evolutionary process. SAWYER *et al.* (1987) first suggested that comparisons of the frequency distributions of classes of mutations can reveal differences in their effect on fitness. Positive selection elevates the likelihood that a change will increase in frequency within a population, whereas negative selection decreases the likelihood for changes to spread (FISHER 1930; WRIGHT 1938). The frequency distribution of amino acid mutations is skewed toward rare variants compared with that of silent mutations at the *gnd* locus in *Escherichia coli* (SAWYER *et al.* 1987). This difference can be attributed to differences in the fitness effects of the two classes of mutations; amino acid changes are, on average, more deleterious than silent mutations in this gene. Comparisons between categories of interspersed mutations control for differences in the evolutionary histories of regions within the sampled data. Interpretations of the statistical analyses do not require an assumption of stationarity or of linkage equilibrium; the action of natural selection can be distinguished from the effects of population structure and history (SAWYER *et al.* 1987).

MCDONALD and KREITMAN (1991) used the same reasoning to infer selection from contrasts between ratios of the number of segregating sites to the number of fixed differences between species, referred to as r_{pd}. r_{pd} decreases monotonically as $N_e s$ varies from negative to positive infinity (AKASHI 1995). Significantly lower r_{pd}'s for replacement changes suggest larger (less negative) selection coefficients for amino acid changes than for silent mutations at the *Adh* locus in Drosophila (MCDONALD and KREITMAN 1991).

Although comparisons of frequency distributions and ratios of polymorphism to divergence can establish the role of natural selection in molecular evolution, such contrasts may not shed light on the sign or magnitude of fitness effects. Comparisons of the evolutionary dynamics of mutations in DNA with theoretical predictions (usually the equilibrium neutral model) allow estimation of parameters such as migration, mutation, and recombination rates, effective population size, and selection intensity (HUDSON 1987; HUDSON *et al.* 1992; FU and LI 1993; FU 1994). In particular, the magnitude of $N_e s$ can be estimated from both frequency data and r_{pd}'s under a model of unidirectional selection, free

Corresponding author: Hiroshi Akashi, Section of Evolution and Ecology, 2320 Storer Hall, University of California at Davis, Davis, CA 95616. E-mail: hakashi@ucdavis.edu

recombination, and stationarity (SAWYER and HARTL 1992; HARTL *et al.* 1994; SAWYER 1997).

The biology of codon usage motivates an evolutionary model that makes both qualitative predictions for the existence of fitness classes of DNA changes and quantitative predictions for the magnitude of selection intensity at silent sites (AKASHI 1995). *E. coli* and *Saccharomyces cerevisiae* exhibit specific patterns of codon usage (reviewed in ANDERSSON and KURLAND 1990). Synonymous codon usage in these organisms is biased toward major codons that generally encode the most abundant tRNA(s) for each amino acid (IKEMURA 1981, 1982; BENNETZEN and HALL 1982; GROSJEAN and FIERS 1982). The degree to which codon usage is biased varies among genes and correlates strongly with protein abundance (BENNETZEN and HALL 1982; GUOY and GAUTIER 1982; IKEMURA 1985). Finally, silent DNA divergence is inversely related to codon usage bias (SHARP and LI 1987; but see also EYRE-WALKER and BULMER 1995). Patterns of codon usage and silent DNA evolution in *D. melanogaster* appear to be similar to those found in *E. coli* and yeast (SHIELDS *et al.* 1988; SHARP and LI 1989).

These patterns suggest a model of "major codon preference," a form of mutation-selection balance at silent sites (SHARP and LI 1986; LI 1987; BULMER 1988). During polypeptide chain elongation in *E. coli,* the arrival time of a cognate tRNA is inversely proportional to its abundance (VARENNE *et al.* 1984; CURRAN and YARUS 1989). Major codons may confer fitness benefits by enhancing translational elongation rates, by lowering the energetic cost of proofreading (rejecting noncognate tRNAs) or by reducing the rate of amino acid misincorporation (BULMER 1988). Because the fitness effect of a single codon is a function of the number of times it is translated, selection intensity for codon bias will be stronger in highly expressed genes. Under the major codon preference model, natural selection favors translationally superior codons for each amino acid, whereas mutation pressure and genetic drift allow nonmajor codons to persist. Codon bias is maintained by positive selection acting on synonymous mutations to major codons (preferred changes) and selection against deleterious (unpreferred) mutations in the opposite direction. Mutational models of codon bias and other selection models do not predict differences in the fitness effects of these categories of mutations.

Mutation-selection-drift at silent sites can be tested by contrasting the evolutionary dynamics of synonymous DNA mutations. In a previous study, AKASHI (1995) showed higher ratios of polymorphism to divergence for preferred than for unpreferred mutations, suggesting a role of selection at silent sites in Drosophila. Maximum likelihood estimates of selection intensity from r_{pd} data appear consistent with a balance among mutation, weak selection, and genetic drift maintaining codon bias in *D. simulans* (AKASHI 1995). Here, we examine a second population genetic prediction of major codon preference; advantageous preferred mutations should segregate at higher frequencies within a population than deleterious unpreferred changes. We infer the action of natural selection and estimate the intensity of selection at silent sites from the frequency distributions of DNA polymorphism in Drosophila.

METHODS AND RESULTS

Theoretical predictions for the effect of natural selection on frequency spectra: SAWYER and HARTL (1992) model the flux of irreversible mutations through a population as a Poisson random field. Their model assumes completely independent dynamics of mutations and a stationary frequency distribution in a haploid population. Let q represent the frequencies of nonancestral nucleotides at polymorphic sites. The expected number of sites at which r nonancestral nucleotides will be observed in a sample of n sequences is a Poisson random variable with mean (HARTL *et al.* 1994)

$$M(r{:}\mu,\gamma) = 2\mu \int_0^1 \frac{1 - e^{-2\gamma(1-q)}}{1 - e^{-2\gamma}} \binom{n}{r} \times q^r (1 - q)^{n-r} \frac{dq}{q(1 - q)}, \quad (1)$$

where μ is the mutation rate (scaled to the effective population size and the number of sites in the region), and γ represents the product of selection coefficient and effective population size, $N_e s$. Both the sign and magnitude of γ are assumed to be constant among sites and over evolutionary time. n_r represents the observed number of sites at which r nonancestral nucleotides and $n - r$ ancestral nucleotides segregate in the sample. Equation (1) also applies to diploid organisms under semidominant fitness effects by letting $\gamma = 2N_e s$.

Equation 1 was solved numerically for the expected frequencies for newly arising polymorphism under negative and positive selection. Figure 1a shows the skew toward rare variants expected under negative selection. As $N_e s$ becomes large and negative, all variants are expected to be "singletons" (*i.e.*, observed in a single sequence in the sample). As $N_e s$ increases (toward positive infinity), the frequency spectrum asymptotically approaches a symmetrical distribution (Figure 1b; in the diploid case, the symmetrical distribution only holds for semidominant fitness effects, $h = 0.5$). Between the two asymptotic values, the frequency spectrum shows the greatest change near zero, especially for weak, negative selection (Figure 2). These patterns suggest that even small fitness differences between preferred and unpreferred mutations could differentiate the frequency distributions of mutations segregating within a species.

Major codons in Drosophila: In Drosophila, both regional mutational biases and variation in selection intensity appear to contribute to variation in synonymous codon usage bias. Correlations between intron and silent-position base composition suggest (assuming intron sequences reflect equilibrium mutational biases) that regional mutational biases explain ~10% of codon bias variation in *D. melanogaster* (KLIMAN and HEY 1994). However, the GC content of coding regions is higher than that of associated introns in 154 of 155 genes in KLIMAN and HEY's study; selection for codon bias may play a prominent role in determining base composition at silent sites, even in low-codon-bias genes.

If differences in $N_e s$ account for patterns of codon usage bias, then codons favored by selection can be identified by comparing codon bias among genes. Under major codon preference, selection intensity at silent sites is stronger in highly expressed genes than in low-expression loci. Major co-

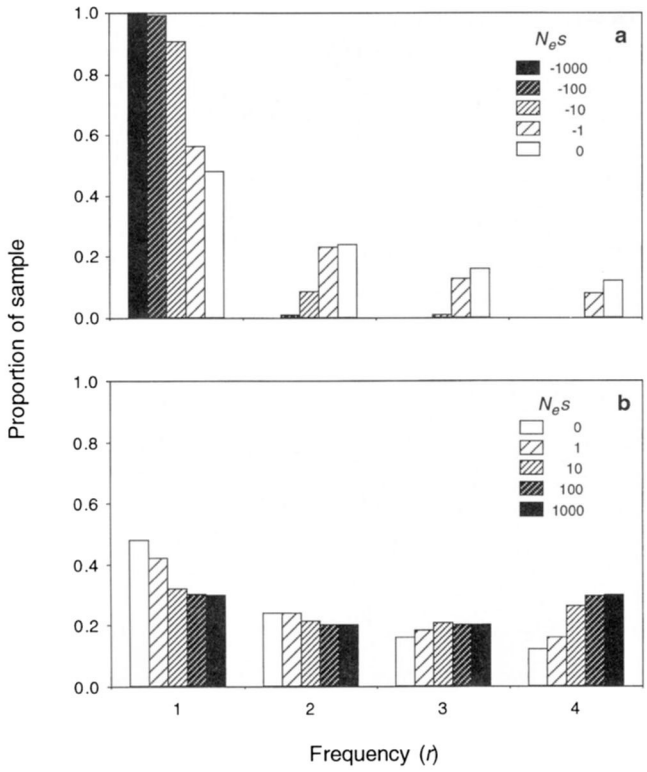

FIGURE 1.—Natural selection and frequency distributions. The effect of negative (a) and positive (b) selection on the proportion of a sample of newly arising polymorphisms at frequency classes one to four in a sample of five sequences (from Equation 1).

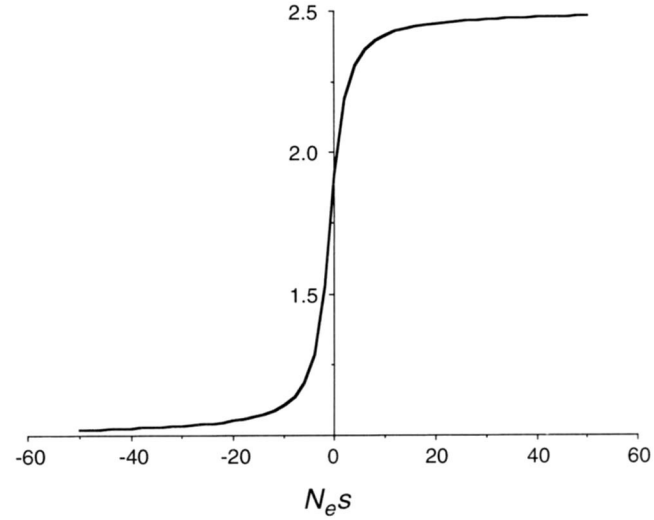

FIGURE 2.—Natural selection and the expected frequency of segregating mutations. The effect of natural selection on the expected frequency of a newly arising mutation in a sample of five sequences (from equation 1).

dons can be identified as those whose frequencies (within a synonymous family) show positive correlations with the degree of bias at other codons in the same gene.

Major codons were identified in *D. melanogaster* in a study of 575 protein coding genes (such codons are referred to as "preferred" codons in AKASHI 1995). To determine whether *D. simulans* and *D. pseudoobscura* share the same major codons as *D. melanogaster*, similar analyses were applied to a smaller number of genes: 29 in *D. simulans* and 22 in *D. pseudoobscura* (Table 1). Because the number of loci is much smaller than in the *D. melanogaster* sample, the genes from *D. simulans* and *D. pseudoobscura* were divided into two categories (low bias and high bias) and codon usage for each synonymous family was compared between the two groups.

The "scaled" chi square (SHIELDS *et al.* 1988) was used to measure the level of bias in a given gene. Chi square values for deviations from an A+T content of 60%, the average base content of *D. melanogaster* introns (SHIELDS *et al.* 1988; CARULLI *et al.* 1993; MORIYAMA and HARTL 1993) were calculated for each synonymous family. The A + T content of *D. simulans* (60.7% A + T for 46 introns) and *D. pseudoobscura* (60.6% A + T for 47 introns) are close to this value. The sum of the 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%) (AKASHI 1995). Among 575 *D. melanogaster* genes, this measure ranges from 0.04 to 1.9 (H. AKASHI, unpublished data).

The *D. simulans* and *D. pseudoobscura* genes were divided into high bias and low bias groups with a cutoff of the scaled chi square (AT60%) of 0.82 for *D. simulans* and 0.7 in *D. pseudoobscura*. The cutoff values were chosen to divide the data into roughly equal numbers of codons in high and low bias groups for each species. A subset of codons identified as major codons in *D. melanogaster* show significantly higher frequencies in the high bias than in the low bias group in both *D. simulans* (19 of 22) and *D. pseudoobscura* (16 of 22). No other codons show a significant increase in codon bias in these species (Table 2). The similarity of codon bias patterns in *D. melanogaster* and *D. simulans* is not surprising given the lack of sequence divergence ($k_s \approx 0.1$ in low bias genes) between these sibling species; differences in major codons are difficult to detect in such closely related taxa. Parallels in codon usage between the more distantly related species, *D. melanogaster* and *D. pseudoobscura*, are more informative. Because both comparisons show no evidence for differences in codon usage; major codons in *D. simulans*, and *D. pseudoobscura* will be assumed to be identical to those of *D. melanogaster*.

***D. melanogaster* and *D. simulans* sequences:** Nine genes for which multiple alleles have been sequenced in *D. melanogaster* and *D. simulans* and at least one other species outside this clade but within the *D. melanogaster* subgroup were examined. GenBank accession numbers or references for these sequences are: *Adh* (*mel*: M17834-37, M19547, M22210, *sim*: X57362-64, M36581, X00607, *ere*: X54116, *ore*: M37837, *yak*-13: X54120, X57365-76, *tei*: X54118). *Adhr* [*mel*: (KREITMAN and HUDSON 1991), *sim*: (SUMNER 1991), *ere*: X54116, *yak*: X54120, *tei*: X54118]. *Amy* (*mel*: L22726-31, L22733, L22735, *sim*-2: D17733-34, *ere*-2: D17727-28, *ore*-2: D21128-29, *yak*-2: D17737-38, *tei*-2: D17735-36). *boss* [*mel*, *sim*, *tei*-3, *yak*-4 (AYALA and HARTL 1993)]. *Mlc1* (*mel*: L37312-17, *sim*: L49010-14, *tei*: L49008, *yak*: L49007). *per* (*mel*: L07817-19, L07821, L07823, L07825, *sim*: L07828-32, *yak*: X61127). *Pgi* (*mel*: L27544-46, L27553-55, *sim*: L27547-51, U20556-59, U20564-65, *yak*-13: L27673-85, *tei*-1: J. H. MCDONALD, personal communication). *Rh3* [*mel*, *sim*, *tei*-5, *yak*-5: (AYALA *et al.* 1993)]. *Zw* (*mel*: U43167, U43165, U42748, U42742, U42738-39, *sim*: L13876, L13878, L13881, L13883, L13891, *yak*: U42750). *mel*, *sim*, *yak*, *tei*, *ere*, and *ore* refer to *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. teisseri*, *D. erecta*, and *D. orena*, respectively. The number of alleles examined for each gene in each species is shown. Six alleles were examined at each of the nine genes in *D. melanogaster*, and five alleles were examined in eight genes in *D. simulans*. All samples are unbiased with respect to allozyme polymorphism except the sequences of *Adh* and *Amy* in *D. melanogaster*.

***D. pseudoobscura* sequences:** The *Adh* region is composed of two structural genes, *Adh* and *Adhr*. We examined 99 se-

TABLE 1

***D. simulans* and *D. pseudoobscura* genes**

D. simulans				*D. pseudoobscura*			
Locus	Codons	Codon bias	Reference	Locus	Codons	Codon bias	Reference
ac	201	0.39	X62400	*ade3*	1364	0.55	X06285
Acp26Aa	255	0.21	X70899	*Adh*	254	1.18	X64489
Acp26Ab	90	0.69	X70899	*Adhr*	278	0.84	X64489
Act88F	376	1.64	M87274	*Amy*	494	1.69	X76240
Adh	256	1.57	X57361	*bcd*	538	0.47	X55735
Adhr	272	0.34	SUMNER (1991)	*Est-5a*	545	0.27	X89085
Amy-d	494	1.70	D17733	*exu2*	497	0.50	L22554
ase	355	0.23	J. HEY[a]	*Gld*	612	0.87	M29299
bcd	29	1.47	M32123	*Hsp83*	375	1.09	X03812
boss	521	0.62	AYALA and HARTL (1993)	*l(2)gl*	1166	0.09	X73259
ci	314	0.25	A. BERRY[a]	*M(3)99D*	134	1.06	S59382
cta	322	0.14	M. WAYNE[a]	*Pcp*	192	0.68	X06285
Est-6	542	0.32	L34263	*per*	1241	0.65	X13878
eve	58	1.20	U32092	*Rh2*	381	1.12	X65878
GstD1	200	2.14	M84577	*Rh3*	382	0.80	X65879
Hsc70-1	104	1.14	J01088	*run*	561	1.13	U22357
Hsp83	375	1.30	X03811	*ry*	1342	0.84	M33977
Mlc1	168	1.05	L08051	*Sry-alpha*	549	0.28	L19536
MtnA	40	1.49	M55407	*Sry-beta*	140	0.30	S77099
per	559	1.18	L07829	*Tl*	1100	1.16	L25390
Pgd	481	1.09	U02288	*Ubx*	258	1.45	X05179
Pgi	558	1.21	L27550	*Uro*	346	0.90	X57113
pn	365	0.47	SIMMONS *et al.* (1994)				
ref(2)P	599	0.40	U23930				
Rh3	383	0.85	AYALA *et al.* (1993)				
sala	139	0.40	M21227				
Sgs3	28	1.04	MARTIN *et al.* (1988)				
Sod	153	1.10	X15685				
su(f)	733	0.37	L09193				
tra	184	0.38	X66930				
v	379	0.82	U27204				
Yp2	348	1.39	L14426				
z	268	0.84	L13049				
Zw	518	1.59	L13876				

Gene symbols are from FlyBase (1995). Number of codons examined (excluding termination codons), codon bias "scaled" chi square (AT60%), and GenBank accession number or references are given for each gene. Only genes for which ≥100 codons have been sequenced were included in the analyses.

[a] Personal communication.

quences of the *D. pseudoobscura Adh* region from flies collected in North America (SCHAEFFER and MILLER 1993). This sample is also random with respect to knowledge of allozyme polymorphism. Sequences from flies collected from Bogota, Columbia, were not included because the *Adh* regions show evidence for divergence from North American populations (SCHAEFFER and MILLER 1991, 1992). Population locations and strain names are given in SCHAEFFER and MILLER (1993). EMBL/GenBank Data Library accession numbers for the 99 sequences are: X62181-238, X64468-89, M60979-88, Y00602, X68159-66. A single *D. miranda* sequence (M60998) was used as an outgroup sequence for these data.

Inferring the direction of synonymous DNA changes: Parsimony assumptions and outgroup sequences were used to infer the direction of synonymous polymorphism in *D. melanogaster* and *D. simulans*. At a given polymorphic site in *D. melanogaster* or *simulans*, mutations were assigned to minimize the number of changes in the phylogenetic tree. Sites at which multiple trees give the fewest number of changes were not included in the analyses. APPENDIX A gives an example of the classification and frequencies of polymorphic and fixed (since the split from the common ancestor to *D. melanogaster*) silent DNA changes among five alleles of the *D. simulans Adhr* locus. All gene sequences and raw data examined in these analyses are available from the authors.

We used a single outgroup, *D. miranda*, to determine the direction of synonymous changes segregating in *D. pseudoobscura*. The number of fixed differences between the *D. pseudoobscura* sequences and the single *D. miranda* sequence are small: five replacement and five silent substitutions at *Adh* and two replacement and four silent substitutions at *Adhr*. At a given polymorphic site within the *D. pseudoobscura* sequences, the nucleotide encoded in the *D. miranda* sequence is considered the ancestral state. Four polymorphic sites (two at *Adh* and two at *Adhr*) at which an ancestral state could not be inferred are not included in the analysis. At these sites, the *D. miranda* sequence does not encode either of the nucleotides segregating within *D. pseudoobscura*.

Comparison of the frequency spectra of synonymous DNA changes: Table 3 and Figure 3 show the frequencies of preferred and unpreferred synonymous DNA changes in *D. melanogaster*, *D. simulans*, and *D. pseudoobscura* (the raw data are

TABLE 2

Major codons in *D. simulans* and *D. pseudoobscura*

aa[a]	*D. simulans*			*D. pseudoobscura*		
Codon	Low	High	*G*	Low	High	*G*
Phe						
TTT	105	29		113	65	
TTC[b]	106	200	74.1*	96	204	45.2*
Leu						
TTA	36	3		28	4	
TTG	107	33		120	56	
CTT	50	19		57	20	
CTC[b]	50	76	10.4	106	130	4.3
CTA	48	12		57	20	
CTG[b]	174	270	69.2*	215	332	56.9*
Ser						
TCT	52	12		74	17	
TCC[b]	80	140	29.8*	86	136	25.9*
TCA	34	11		51	17	
TCG[b]	75	79	0.128	107	120	3.9
Tyr						
TAT	77	25		72	50	
TAC[b]	89	151	43.5*	97	171	17.7*
Cys						
TGT	28	8		34	25	
TGC[b]	52	73	15.1*	64	102	6.3
Pro						
CCT	52	13		43	21	
CCC[b]	69	134	65.0*	134	179	13.7*
CCA	90	22		90	38	
CCG	66	53	0.0	85	106	3.9
His						
CAT	65	33		95	60	
CAC[b]	73	93	12.4*	86	118	12.9*
Gln						
CAA	110	17		96	24	
CAG[b]	172	211	74.0*	203	207	37.4*
Arg						
CGT	62	42		68	58	
CGC[b]	92	118	34.4*	103	137	6.2
CGA	44	12		47	24	
CGG	45	26		39	48	1.4
AGA	47	6		23	12	
AGG	30	15		18	29	2.4
Ile						
ATT	103	71		96	96	
ATC[b]	96	222	70.4*	116	234	37.5*
ATA	64	19		73	31	
Thr						
ACT	68	28		73	46	
ACC[b]	91	179	37.6*	118	186	19.4*
ACA	59	19		91	41	
ACG	55	79	2.7	96	126	3.6
Asn						
AAT	137	44		142	110	
AAC[b]	125	199	66.3*	170	243	14.5*
Lys						
AAA	114	23		103	29	
AAG[b]	149	312	115.7*	214	280	52.5*
Ser						
AGT	78	14		82	37	
AGC[b]	91	105	41.1*	136	152	16.2*
Val						
GTT	77	33		86	22	
GTC[b]	74	124	15.2*	103	135	3.6
GTA	34	7		41	25	
GTG[b]	138	173	4.9	164	237	18.2*
Ala						
GCT	109	57		111	76	
GCC[b]	129	263	56.1*	242	336	17.9*
GCA	62	19		87	52	
GCG	53	75	1.3	83	105	1.3
Asp						
GAT	159	97		172	149	
GAC[b]	129	200	30.4*	155	174	2.7
Glu						
GAA	139	39		132	40	
GAG[b]	198	310	84.0*	215	288	61.8*
Gly						
GGT	67	65		83	66	
GGC[b]	86	242	72.4*	208	293	29.8*
GGA	122	82		124	75	
GGG	27	11		60	43	

[a] aa, amino acid.

[b] A major codon identified in *D. melanogaster* (AKASHI 1995). The *G* test statistic (1 d.f.) for heterogeneity within a synonymous family is shown for codons that increase in frequency in high bias genes. Serine is divided into a twofold and fourfold family.

* Codons showing a significant relationship ($P < 0.05$) using the sequential Bonferroni test (RICE 1989) are defined as major codons.

shown in APPENDIXES B and C). Since no prediction is made for the fitness effect of mutations occurring within classes, major → major and nonmajor → nonmajor changes, are not analyzed. Four codons in the *D. simulans* data and five codons in *D. pseudoobscura* show multiple mutations at the same site. In both species, three such sites show two newly arising mutations to the same category (major, nonmajor, or replacement). At such codons, a single mutation between the ancestral codon and the new state is assumed to be segregating at the sum of the frequencies of the two mutations. At one codon in *D. simulans* and two codons in *D. pseudoobscura*, two silent mutations occur at the same site and one of the mutations falls within a bias class. At these codons, the new mutation predicted to have a fitness effect (the one occurring between classes) is considered to be the only mutation segregating. This classification scheme does not bias the analysis; elimination of multiply mutated codons has little effect on the results presented below.

In *D. melanogaster*, only three preferred silent mutations were found segregating among six alleles of each of nine genes. No comparisons of frequency spectra could be made in this species. In *D. simulans*, the 24 preferred polymorphisms are segregating at higher frequencies than the 87 unpreferred polymorphisms (Mann-Whitney test, $z = 1.71$, $P = 0.044$, one-tailed). Similarly in *D. pseudoobscura*, the 24 preferred mutations are segregating at significantly higher fre-

TABLE 3

Frequency spectra of DNA polymorphism in *D. melanogaster*, *D. simulans* and *D. pseudoobscura*

Species	r	n_r Unpref	Pref	Rep
D. simulans (n = 5)	1	55	12	9
	2	22	4	1
	3	7	5	0
	4	3	3	0
D. melanogaster (n = 6)	1	26	2	10
	2	12	0	3
	3	14	0	2
	4	7	1	0
	5	10	0	2
D. pseudoobscura (n = 99)	1–2	51	12	4
	3–8	10	3	1
	9–28	4	6	2
	29–98	4	3	2

The number of nonancestral mutations, n_r, segregating at frequency r in a sample of n sequences are shown for unpreferred (Unpref) and preferred (Pref) silent changes, and replacement (Rep) mutations. Raw data are shown in Appendixes B and C.

quencies than the 69 unpreferred changes (z = 1.72, P = 0.043). Combining the independent results from *D. simulans* and *D. pseudoobscura* (FISHER 1954) strengthens the statistical evidence for differences in frequency spectra (χ^2 = 10.9, d.f. = 4, P = 0.014).

In *D. simulans* and *D. melanogaster*, the frequency distributions of replacement changes are significantly skewed toward rares compared with that of the pooled silent changes, suggesting that amino acid changes have stronger deleterious effects than silent mutations (AKASHI 1996) (data for *Adh* and *Amy* were not included in *D. melanogaster* because the sequences were not examined randomly with respect to amino acid polymorphism). The frequencies of the nine amino acid changes in the *D. pseudoobscura Adh* region do not differ significantly from that of the 69 unpreferred synonymous mutations (z = 1.82, P = 0.07, two-tailed) or from that of the 24 preferred changes (z = 0.81, P = 0.42) or from the pooled silent changes (z = 1.61, P = 0.11); the null hypothesis of equal fitness effects cannot be rejected between amino acid and silent mutations. These probabilities are calculated as two-tailed because no *a priori* prediction was made for the fitness effect of replacement changes relative to that of silent mutations. Frequency data for a larger number of Drosophila loci will be required to determine if segregating amino acid mutations are, on average, more strongly selected against than silent mutations as SAWYER *et al.* (1987) found at the *E. coli gnd* locus.

Estimating selection intensity at silent sites: Since each n_r in Equation (1) is an independent Poisson random variable, the likelihood of the distribution, $\{n_r\}$, in the sample is (HARTL *et al.* 1994)

$$L(\{n_r\};\mu,\gamma) = \prod_{r=1}^{n-1} e^{-M(r;\mu,\gamma)} \frac{M(r;\mu,\gamma)^{n_r}}{n_r!}.$$

$\hat{\mu}_{MLE}$ can be expressed as a function of (SAWYER 1994)

$$\hat{\mu}_{MLE} = \frac{n_{tot}}{G_{tot}}$$

where

$$G(r,\gamma) = \frac{M(r;\mu,\gamma)}{2\mu},$$

$$n_{tot} = \sum_{r=1}^{n-1} n_r,$$

and

$$G_{tot} = \sum_{r=1}^{n-1} G(r,\gamma).$$

Thus, finding $\hat{\gamma}_{MLE}$ can be reduced to the one-dimensional maximization (SAWYER 1997)

$$L(\gamma) = L[\{n_r\};\hat{\mu}_{MLE}(\gamma),\gamma] = C(\{n_r\}) \prod_{r=1}^{n-1} \left(\frac{G(r,\gamma)}{G_{tot}}\right)^{n_r} \quad (2)$$

where $C(\{n_r\})$ is a constant.

Numerical solutions to Equation (2) using the data in Table 3 give MLE $N_e s$ values shown in Table 4. Mutations segregating among the 99 *D. pseudoobscura* alleles were divided into

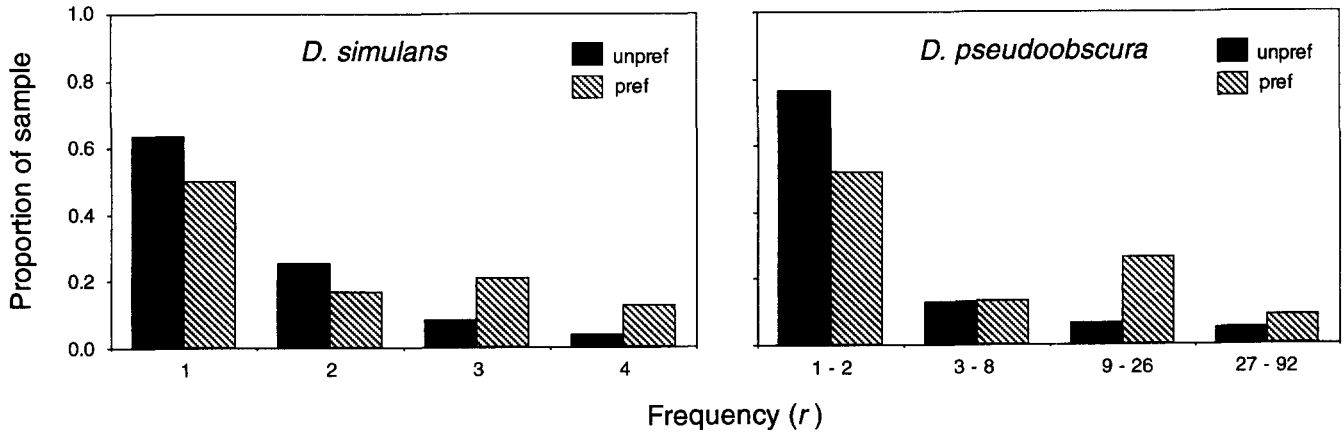

FIGURE 3.—Frequency distributions of preferred and unpreferred synonymous DNA mutations in Drosophila. The proportion of unpreferred (black) and preferred (striped) mutations segregating at the given frequencies are shown. (a). Pooled data from eight *D. simulans* genes. (b) Data from the *D. pseudoobscura Adh* and *Adhr* genes. The frequencies (r = 1–98) were divided into four intervals, each with $n_r/n_{tot} \approx 0.25$ for neutral mutations under the SAWYER-HARTL model (data from Table 3).

TABLE 4

Selection intensity at silent sites in Drosophila

Species	Loci	Analyses	Class	$N_e s$	95% CI
D. simulans	8	fd	Unpreferred	−2.1	−3.2, −1.0
			Preferred	0.2	−2.0, 13.5
		r_{pd}	Unpreferred	−1.2	−2.1, −0.3
			Preferred	0.6	−0.7, 3.2
D. pseudoobscura	*Adh* + *Adhr*	fd	Unpreferred	−4.6	−12.1, −2.4
			Preferred	−1.4	−4.3, 0.3

The loci examined in *D. simulans* and *D. melanogaster* are listed in the text. Analyses, method of estimating $N_e s$; fd, frequency distributions; and r_{pd}, ratios of polymorphism to divergence. The small number of replacement mutations in these species preclude estimation of selection intensity.

four frequency classes (see Figure 3) so that confidence intervals could be obtained by bootstrapping n_r values. Because all mutations within a given class are pooled, these estimates of $N_e s$ reflect a weighted average for each category and do not address within-class variation in fitness effects. Also, since the SAWYER-HARTL method is limited to estimating selection intensity for mutations which contribute to within-species polymorphism; sites under strong negative selection are, *a priori*, not considered in the analyses. MLE $N_e s$ values were also calculated from the ratio of polymorphism to divergence in *D. simulans* for r_{pd}'s of 87/13 and 24/13 for unpreferred and preferred mutations, respectively (data for the eight genes were combined). The parameter t_{div} was estimated from the ratio of polymorphism to divergence for the introns of the eight genes in *D. simulans*, 66/25. The method followed that of AKASHI (1995).

Confidence intervals for these $N_e s$ values were estimated using parametric bootstrap simulations (EFRON and TIBSHIRANI 1993, p. 54). For each observed frequency distribution, $\{n_r\}$, bootstrap frequency distributions were generated by resampling each of the observed n_r values from a Poisson distribution of mean n_r. For r_{pd} data, the numbers of polymorphic and fixed differences for both the intron and silent classes were resampled from Poisson distributions with the observed values as means. MLE $N_e s$ values were determined numerically for 10,000 iterations. This method gives 95% confidence intervals shown in Table 4. Under the assumptions of the Poisson random field model, these confidence intervals include both evolutionary (stochastic) and sampling variance. All $N_e s$ estimates fell within the mid-10% interval of the bootstrap distribution. The larger confidence interval for unpreferred mutations in *D. pseudoobscura*, despite larger sample sizes than in *D. simulans*, reflects a weaker fit of the observed frequency distribution to the expected distribution under the MLE $N_e s$. Similarly, the large confidence interval for preferred changes in *D. simulans* reflects the combined effect of a small sample size, poor fit to the expected distribution, and the lack of change in frequency distributions under positive selection (Figure 2).

DISCUSSION

Inferring natural selection from frequency spectra: Attempts to infer the action of natural selection from frequency spectra of naturally occurring DNA and protein polymorphism have been hindered by comparing observed data to a neutral equilibrium expectation (EWENS 1977). The methods of OHTA (1975), EWENS-WATTERSON (EWENS 1972; WATTERSON 1977), and TAJIMA (1989a) assume that mutations segregating within a population have reached a stationary frequency distribution. Under a constant mutation rate and effective population size, the number of segregating sites and the frequency distribution of neutral mutations will reach a steady-state over time. At stationarity, the frequencies of individual mutations change within a population (many are lost and a few go to fixation) but the frequency spectrum remains relatively constant. Statistically significant excesses of rare allozyme alleles in Drosophila and humans are consistent with weakly deleterious amino acid changes in genetic regions at stationarity (LATTER 1975; OHTA 1975; WATTERSON 1977; CHAKRABORTY *et al.* 1980; KEITH *et al.* 1985). However, neutral variation in the approach to equilibrium (*i.e.*, during population expansion following a bottleneck) can produce similar frequency distributions (NEI *et al.* 1975; GRIFFITHS 1979; WATTERSON 1983; MARUYAMA and FUERST 1984).

To distinguish between the effects of natural selection and evolutionary history, TAJIMA (1989a) suggested comparison of separate classes of closely linked DNA mutations to the neutral equilibrium expectation. Selection can be inferred if only one of the categories of mutations departs from this expectation. TAJIMA found that, in *D. melanogaster*, large insertions (presumably transposable element insertions) segregate at significantly lower frequencies than that expected for neutral changes at equilibrium. Restriction site polymorphisms, however, conform to this expectation. Although these findings support the action of selection against transposon insertions, the method is limited by the requirement for a neutral class of changes near equilibrium. Since the approach to stationarity is extremely slow (TAJIMA 1989b), stationary distributions may be uncommon in nature. In addition, the statistical power of each analysis must be examined to determine whether the inability to reject the null hypothesis of the Tajima test (for presumably neutral mutations) can be interpreted as evidence that a region is at equilibrium.

Directly comparing the frequency distributions of classes of mutations controls for the evolutionary history of the sample of sequences. If the classes of muta-

tions are randomly interspersed within a genetic region, departures from stationarity will have an equivalent impact on their frequency distributions; differences in their frequency spectra can be attributed to natural selection (SAWYER *et al.* 1987). The power of the analysis of frequency spectra can be enhanced by using phylogenetic inference to determine the direction of newly arising mutations. In the absence of such knowledge, the numbers of mutations segregating at frequencies r and $(n - r)$ must be pooled. However, this increase in statistical power could be lessened by misidentification of ancestral states in DNA.

In the present analyses, violation of parsimony assumptions could bias ancestral state reconstruction and lead to errors in categorizing mutations. PERNA and KOCHER (1995) pointed out that unequal forward and backward mutation rates can lead to errors in ancestral state reconstruction in mtDNA. Similarly, if fixation probabilities or mutation rates are higher for preferred than for unpreferred changes, then the likelihood of two preferred substitutions can be higher than that for a single unpreferred fixation. Inferring ancestral states from the tree with the fewest number of changes can lead to under- and overestimating the numbers of preferred and unpreferred changes, respectively. The magnitude of this bias depends on the product of fixation probabilities, mutation rates, and the number of generations separating the observed sequences.

In the statistical analyses of population genetic predictions, biased ancestral state reconstruction (BASR) will not cause a departure from the null hypothesis in the direction predicted by major codon preference. When examining within and between species data, fixed differences will be more prone to misclassification than polymorphic changes because the direction of segregating mutations are generally inferred using more data. BASR will decrease r_{pd} for unpreferred changes and increase it for preferred mutations. Major codon preference, however, predicts *higher* r_{pd}'s for unpreferred than for preferred mutations. In the present study, parallel mutations occurring within *D. pseudoobscura* and between *D. pseudoobscura* and *D. miranda* could cause some silent changes to be erroneously categorized (the low numbers of fixed differences between *D. pseudoobscura* and *D. miranda* suggest that double mutations are rare in these data). In the absence of fitness differences between the classes, mixing categories of mutations will not cause a departure from equivalent frequency distributions. Under major codon preference, the statistical test becomes a more conservative one because BASR will result in more similar frequency distributions.

Comparisons of the frequency spectra of classes of synonymous DNA changes also distinguish between major codon preference and mutational models of codon bias (AKASHI 1995). Directional mutation pressure can bias equilibrium base composition (SUEOKA 1988). In Drosophila, higher mutation rates in the direction A/T → C/G than in the reverse direction could explain codon usage bias. These classes of mutations generally correspond to preferred and unpreferred mutations under major codon preference. Mutational biases, in the absence of natural selection, predict higher *numbers* of segregating mutations (per site) for preferred than for unpreferred changes but will not affect the frequency spectra of the two classes of mutations. Recent *changes* in per locus mutation rates, however, could alter frequency spectra. If the ratio of mutation rates to unpreferred and preferred changes has increased since the most recent common ancestor of the alleles segregating within a species, then the frequency spectra of unpreferred mutations will be skewed toward rares. In the present context, it is unlikely that such changes in mutation pressure have occurred independently in both *D. simulans* and *D. pseudoobscura*.

Neither departures from stationarity and linkage equilibrium, errors in classifying silent changes, nor mutational biases are likely to explain the observed difference between the frequency spectra of preferred and unpreferred silent changes in *D. simulans* and *D. pseudoobscura*. These analyses appear to confirm the existence of (at least) two fitness classes of silent mutations in Drosophila.

Intensity of selection for codon bias: The major codon preference model requires a restricted range of selection intensity at silent sites. If $N_e s$, the product of selection coefficient and effective population size, falls much below unity, then genetic drift will overwhelm selection and codon bias will approach mutational equilibrium. Differences in the frequency spectra of preferred and unpreferred mutations confirm the action of natural selection on synonymous changes ($|N_e s| > 0$). The persistence of nonmajor codons through mutation pressure and genetic drift, however, also requires an upper limit on selection intensity at silent sites. If $N_e s$ becomes greater than three or four (depending on mutational biases), all sites will be fixed for major codons. Because weak selection ($s \approx 1/N_e$) for many organisms is several orders of magnitude smaller than that which can be measured in the laboratory or in nature, analyses of DNA sequence variation may provide the only means of estimating the strength of selection at silent sites.

SAWYER and HARTL's (1992) Poisson random field (PRF) theory gives rise to two methods for estimating $N_e s$. The two methods employ different aspects of within and between species molecular variation. Fitting a likelihood function to observed frequencies of mutations does not employ information concerning between-species divergence (outgroup sequences are used only to determine the direction of synonymous changes). In contrast, r_{pd} analyses requires observations of the number of segregating sites within species and the number of fixed differences between species; the frequencies of polymorphic mutations are not considered. Unlike direct comparisons between the evolutionary dynamics of classes of mutations, estimating selection intensity

under PRF theory requires a number of important assumptions about the data: random sampling of sequences from a panmictic population, independent evolution at all sites (free recombination and independent effects on fitness), and a stationary frequency distribution.

Departures from these assumptions can bias estimates of $N_e s$ and/or inflate the variance of such estimates. Sampling from a structured population will elevate both the frequency of rare mutations and the number of segregating sites relative to that expected for a panmictic population. This will bias the PRF method toward negative values of $N_e s$ for both frequency distributions and r_{pd} analyses. The *D. simulans* alleles examined above were sequenced from worldwide population samples. Geographic surveys of allozyme variation suggested little population differentiation in this species except in the Seychelles and in Madagascar (CHOUDHARY and SINGH 1987), but subsequent analyses of nucleotide variation has revealed significant differentiation between Africa and North Carolina, at least for one gene region (BEGUN and AQUADRO 1995). The samples examined here do not include sequences from the Seychelles or from Madagascar, but include flies from Africa, Australia, North America, and the Caribbean; this sampling scheme may have introduced a bias toward negative values of $N_e s$ in *D. simulans*. The North American *D. pseudoobscura* sequences were collected from various locations in North America but show no evidence for population differentiation (SCHAEFFER and MILLER 1992).

Genetic linkage dramatically increases the variance of the number of segregating sites within a population sample (reviewed in HUDSON 1990). Nonindependence among the frequencies of mutations will result in larger confidence intervals for estimates of $N_e s$. The *D. simulans* data are pooled from different genes; rates of recombination should be high between genes (several are located on different chromosomes), but strong linkage exists among sites within a gene. Physical linkage will inflate the variance of $N_e s$ beyond that obtained under PRF assumptions (Table 4). DNA sequence polymorphism among the 99 *D. pseudoobscura Adh* sequences show evidence for high levels of recombination; SCHAEFFER and MILLER (1993) estimate per site recombination rates at ~10-fold greater than mutation rates.

Finally, departures from stationarity will bias estimates of $N_e s$ in opposite directions for frequency spectra and ratios of polymorphism to divergence. In regions of the genome on the approach to equilibrium, the number of segregating sites will be smaller (lower r_{pd}) and the frequency spectra will be skewed toward rare mutations compared with that expected at stationarity. Estimates of $N_e s$ will be biased in the positive and negative directions for r_{pd} and frequency data, respectively (assuming t_{div} reflects the neutral, equilibrium value). For a genetic region in the approach to stationarity from an excess of polymorphism (*i.e.*, following a decrease in effective population size), the number of segregating sites will be higher (higher r_{pd}) and the frequency spectra will be skewed toward common mutations compared with that expected at stationarity. Estimates of $N_e s$ will be biased in the negative and positive directions for r_{pd} and frequency data, respectively (again assuming t_{div} reflects the neutral equilibrium value).

TABLE 5

TAJIMA tests applied to DNA polymorphism at the *D. pseudoobscura Adh* region

Region	No. of sites	S	k	D
5′ flanking	39	6	0.43	−1.40
Adh adult leader	80	10	0.28	−2.15
Adh adult intron	700	129	12.35	−1.61
Adh larval leader	46	2	0.10	−1.14
Adh intron 1	55	20	2.25	−1.18
Adh intron 2	59	25	4.92	0.05
Adh 3′ leader	182	8	0.44	−1.7
intergenic	153	11	0.71	−1.71
Adhr intron 1	205	26	1.66	−1.95
Adhr intron 2	53	18	2.16	−1.05
Adhr 3′ flanking	94	6	0.35	−1.54

The numbers of sites include only positions that do not align with gaps among the 99 alleles of the *D. pseudoobscura Adh* region. The number of segregating sites, S, includes all positions at which at least one mutation is segregating. Sites at which more than two bases are segregating were counted as multiple "segregating sites." Average pairwise differences, k, and the Tajima test statistic, D, were calculated according to TAJIMA (1989). Data available from authors. Note that two of the regions, *Adh* intron 2 and a small region (8 bp) of the *Adh* adult intron show evidence for epistatic selection maintaining linkage disequilibrium (SCHAEFFER and MILLER 1992; KIRBY *et al.* 1995).

Negative values of Tajima's D statistic suggest that stationarity may be violated in the *D. pseudoobscura Adh* region (Table 5). This may explain why the HARTL *et al.* method gives negative selection coefficients for putatively *advantageous* preferred silent changes. The Tajima test, however, has little power to detect departures from stationarity for the small sample sizes ($n = 5$) examined in *D. simulans* (SIMONSEN *et al.* 1995).

Estimating selection intensity using both frequency spectra and ratios of polymorphism to divergence also tests the equilibrium assumption of the Poisson random field method. Overlapping estimates of selection intensity for both preferred and unpreferred mutations in *D. simulans* from frequency spectra and from r_{pd}'s (Table 4) suggest that the equilibrium assumption may be appropriate for these data. The MLE values of $|N_e s| \approx 1$ suggest that natural selection operates at the limit of efficacy in the face of stochastic processes to maintain codon bias in *D. simulans*. However, linkage among sites within genes will inflate the confidence intervals shown in Table 4. Because the confidence intervals on $N_e s$ approach those calculated under free recombination as the number of genes increases (the rate of approach has

not been investigated), stronger evidence for "weak selection" will require overlapping estimates of selection intensity from multiple regions of the genome.

Major codon preference *vs.* conflicting selection pressure in Drosophila: Although differences in the frequency spectra and r_{pd}'s for preferred and unpreferred mutations are consistent with major codon preference, these findings do not exclude the possibility that other selection pressures contribute to patterns of codon bias in Drosophila. Phylogenetic persistence of nonmajor codons suggests site-specific codon bias in prokaryotes (MAYNARD SMITH and SMITH 1996). BULMER (1988) found a reduction in codon bias at the beginning of *E. coli* genes. Reduced synonymous divergence in these regions suggests selection in favor of nonmajor codons perhaps due to constraints of ribosome binding or mRNA secondary structure (EYRE-WALKER and BULMER 1993). Nonmajor codons may also confer fitness benefits (relative to their translationally more efficient counterparts) by reducing the rate of translational elongation. In *E. coli* and *S. cerevisiae*, programmed frameshift events (CURRAN and YARUS 1988; BELCOURT and FARABAUGH 1990; CURRAN 1993) and proper protein folding (PURVIS *et al.* 1987; CROMBIE *et al.* 1992, 1994) can require slowdowns at specific locations on an elongating polypeptide. The extent to which nonmajor codons are maintained by selection, however, is not established.

Conflicting selection pressures could account for the findings presented here if selection is, on average, less intense at codons where the nonmajor codon confers a fitness benefit than at codons where the major codon is favored. Although comparisons of the evolutionary dynamics of categories of mutations can establish differences in the fitness effect of classes of changes in DNA, in the absence of a canonically neutral class of mutations, such analyses cannot reveal the *sign* of selection coefficients. Departures from equal fitness effects can always be explained by differing levels of negative selection. Confirmation of major codon preference as the predominant explanation for the maintenance of codon bias in Drosophila may require evidence for the action of *positive* selection acting on preferred silent changes.

We are grateful to STANLEY SAWYER, MARTY KREITMAN, and JOHN GILLESPIE for many insightful discussions. JOHN MCDONALD kindly provided unpublished DNA sequence data. We also thank BRIAN CHARLESWORTH, NICOLE T. PERNA, ELI STAHL, CHUNG-I WU and two anonymous reviewers for their comments and criticism which helped to improve this manuscript. Differences in the frequency spectra of silent polymorphism in *D. simulans* have been noted independently by RICHARD KLIMAN and ADAM EYRE-WALKER. This research was supported by National Institutes of Health grant GM-42472 to S.W.S. H.A. was a Howard Hughes Medical Institute Predoctoral Fellow, was funded by a postdoctoral fellowship from the Center for Population Biology at the University of California, Davis, and is currently funded by a National Science Foundation/Sloan Foundation Postdoctoral Fellowship in Molecular Evolution.

LITERATURE CITED

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

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

ANDERSSON, S. G. E., and C. G. KURLAND, 1990 Codon preferences in free-living microorganisms. Microbiol. Rev. **54:** 198–210.

AYALA, F. J., B. S. W. CHANG and D. L. HARTL, 1993 Molecular evolution of the *Rh3* gene in Drosophila. Genetica **92:** 23–32.

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.

BEGUN, D. J., and C. F. AQUADRO, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **356:** 519–20.

BEGUN, D. J., and C. F. AQUADRO, 1995 Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140:** 1019–1032.

BELCOURT, M. F., and P. J. FARABAUGH, 1990 Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. Cell **62:** 339–352.

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

BULMER, M., 1988 Are codon usage patterns in unicellular organisms determined by selection-mutation balance. J. Evol. Biol. **1:** 15–26.

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.

CHAKRABORTY, R., P. A. FUERST and M. NEI, 1980 Statistical studies on protein polymorphims in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics **94:** 1039–63.

CHOUDHARY, M., and R. SINGH, 1987 A comprehensive study of genetic variation in natural populations of *Drosophila melanogaster*. III. Variations in genetic structure and their causes between *Drosophila melanogaster* and its sibling species *Drosophila simulans*. Genetics **117:** 697–710.

CROMBIE, T., J. C. SWAFFIELD and A. J. P. BROWN, 1992 Protein folding within the cell is influenced by controlled rates of polypeptide elongation. J. Mol. Biol. **228:** 7–12.

CROMBIE, T., J. P. BOYLE, J. R. COGGINS and A. J. P. BROWN, 1994 The folding of the bifunctional TRP3 protein in yeast is influenced by a translational pause which lies in a region of structural divergence with *Escherichia coli* indoleglycerol-phosphate synthase. Eur. J. Biochem. **226:** 657–664.

CURRAN, J. F., 1993 Analysis of effects of tRNA:message stability on frameshift frequency at the *Escherichia coli* RF2 programmed frameshift site. Nucleic Acids Res. **21:** 1837–1843.

CURRAN, J. F., and M. YARUS, 1988 Use of tRNA suppressors to probe regulation of *Escherichia coli* release factor 2. J. Mol. Biol. **203:** 75–83.

CURRAN, J. F., and M. YARUS, 1989 Rates of aminoacyl-tRNA selection at 29 sense codons *in Vivo*. J. Mol. Biol. **209:** 65–77.

EFRON, B., and R. J. TIBSHIRANI, 1993 *An Introduction to the Bootstrap*. Chapman & Hall, New York.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

EWENS, W. J., 1977 Population genetics theory in relation to the neutralist-selectionist controversy. Adv. Hum. Genet. **8:** 67–134.

EYRE-WALKER, A., and M. BULMER, 1993 Reduced synonymous substitution rate at the start of enterobacteria genes. Nucleic Acids Res. **21:** 4594–4603.

EYRE-WALKER, A., and M. BULMER, 1995 Synonymous substitution rates in enterobacteria. Genetics **140:** 1407–1412.

FISHER, R. A. 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, England.

FISHER, R. A. 1954 *Statistical Methods for Research Workers*. Ed. 12. Oliver & Boyd, Edinburgh.

FLYBASE, 1995 The Drosophila Genetic Database. Available from the ftp.bio.indiana.edu network server and Gopher site.

FU, Y.-X., 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

FU, Y.-X., and W-H. LI, 1993 Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7064.

GRIFFITHS, R. C., 1979 A transition density expansion for a multi-allele diffusion model. Adv. Appl. Prob. **11:** 310–325.

GROSJEAN, H., and W. FIERS, 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., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

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. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

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., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158:** 573–597.

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

KEITH, T. P., L. D. BROOKS, R. C. LEWONTIN, J. C. MARTINEZ-CRUZADO and D. L. RIGBY, 1985 Nearly identical allelic distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. Mol. Biol. Evol. **2:** 206–216.

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

KIMURA, M., and T. OHTA, 1971 Protein polymorphism as a phase of molecular evolution. Nature **229:** 467–469.

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

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.

LATTER, B. D. H., 1975 Enzyme polymorphisms: gene frequency distributions with mutation and selection for optimal activity. Genetics **79:** 325–331.

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

MARTIN, C. H., C. A. MAYEDA and E. M. MEYEROWITZ, 1988 Evolution and expression of the *Sgs-3* glue gene of Drosophila. J. Mol. Biol. **201:** 273–287.

MARUYAMA, T., and P. A. FUERST, 1984 Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics **108:** 745–763.

MAYNARD SMITH, J., and N. H. SMITH, 1996 Site-specific codon bias in bacteria. Genetics **142:** 1037–1043.

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.

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

NEI, M., T. MARUYAMA and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–40.

OHTA, T., 1975 Statistical analyses of *Drosophila* and human protein polymorphisms. Proc. Natl. Acad. Sci. USA **72:** 3194–3196.

PERNA, N. T., and T. D. KOCHER, 1995 Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. **12:** 359–361.

PURVIS, I. J., A. J. E. BETTANY, T. C. SANTIAGO, J. R. COGGINS, K. DUNCAN *et al.*, 1987 The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*: a hypothesis. J. Mol. Biol. **193:** 413–417.

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

SAWYER, S. A., 1997 Estimating selection and mutation rates from a random field model for polymorphic sites, pp. 193–205 in *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, Vol. 87, edited by P. DONNELLY and S. TAVARE. Springer-Verlag, New York.

SAWYER, S. A., D. E. DYKHUIZEN and D. L. HARTL, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA **84:** 6225–6228.

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

SCHAEFFER, S. W., and E. L. MILLER, 1991 Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogata population of *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA **88:** 6097–6101.

SCHAEFFER, S. W., and E. L. MILLER, 1992 Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471–480.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

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.

SIMMONS, G. M., W. KWOK, P. MATULONIS and T. VENKATESH, 1994 Polymorphism and divergence at the *prune* locus in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **11:** 666–671.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

SUEOKA, N., 1988 Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 2653–2657.

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

TAJIMA, F., 1989a Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAJIMA, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

VARENNE, S., J. BUC, R. LLOUBES and C. LAZDUNSKI, 1984 Translation is a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Biol. Chem. **180:** 549–576.

WATTERSON, G. A., 1977 Heterosis or neutrality? Genetics **85:** 789–814.

WATTERSON, G. A., 1983 Allele frequencies after a bottleneck. Statistics Research Report no. 83, Monash Univeristy, Victoria.

WRIGHT, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA **24:** 253–259.

Communicating editor: W-H. LI

APPENDIX A

Polymorphic and fixed differences at the *D. simulans Adhr* gene

Cod	*mel*	*sim*	*tei*	*ere*	Mutation	P/F	Class	n_r
5	ACG	ACA ACG	ACG	ACG	ACG → ACA	P	NN	1
34	CTG	TTG	CTG	CTG	CTG → TTG	F	Unpref	
46	ATC	ATC ATT	ATC	ATC	ATC → ATT	P	Unpref	1
53	AAG	AAA AAG	AAG	AAG	AAG → AAA	P	Unpref	2
56	ACC	ACC ACT	ACC	ACC	ACC → ACT	P	Unpref	1
66	ACC	ACC ACA	ACC	ACC	ACC → ACA	P	Unpref	1
81	GTC	GTC GTA	GTC	GTC	GTC → GTA	P	Unpref	1
119	GTG	GTA GTG	GTG	GTG	GTG → GTA	P	Unpref	1
132	GGG	GGG GGA	GGG	GGG	GGG → GGA	P	NN	1
139	TCG	TCC TCG	TCG	TCG	TCG → TCC	P	MM	1
156	AAA	AAG AAA	AAA	AAG	AAA → AAG	P	Pref	3
176	GGG	GGA	GGG	GGG	GGG → GGA	F	NN	
177	GTA	GTC GTT	GTC	GTG	GTC → GTT	P	Unpref	1
182	GTT	GTT GTG	GTT	GTT	GTT → GTG	P	Pref	1
185	GGT	GGT GGC	GGG	GGT	GGT → GGC	P	Pref	1
186	CCT	CCC CCT	CCT	CCC	CCT → CCC	P	Pref	4
193	CGG	CAG	CGG	CGG	CGG → CAG	F	Rep	
206	GCC	GCC GCT	GCC	—	GCC → GCT	P	Unpref	1
210	CGG	CGT CGA	CGA	—	CGA → CGT	P	NN	1
219	GTT ATT	GTG	GTT	—	GTT → GTG	F	Pref	
224	ATT	ATT ATA	ATT	—	ATT → ATA	P	NN	2
234	GGT	GGC GGT GGG	GGT	—	GGT → GGC	P	Pref	3
					GGT → GGG	P	NN	1
259	TTC	TTC TTT	TTC	—	TTC → TTT	P	Unpref	1

Codon one is the start codon and the numbering follows the coding sequence of KREITMAN and HUDSON 1991. Outgroup sequences are from *D. melanogaster (mel), D. teisseri (tei),* and *D. erecta (ere).* The *D. erecta* sequence is partial. P/F refers to polymorphic (P) or fixed (F) mutations. Class refers to preferred (Pref), unpreferred (Unpref), major-to-major (MM), and nonmajor-to-nonmajor (NN) silent mutations and replacement (Rep) changes. n_r refers to the number of alleles carrying the newly arising mutation.

APPENDIX B

Frequency spectra of DNA polymorphism at the *D. pseudoobscura Adh* and *Adhr* loci

Adh (n = 99)										*Adhr* (n = 99)									
Unpref		Pref		MM		NN		Rep		Unpref		Pref		MM		NN		Rep	
r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r
1	16	1	2	1	1	2	1	98	1	1	20	1	8	1	2	1	4	1	3
2	5	2	1			52	1			2	10	2	1			2	1	2	1
3	1	3	2							3	3	4	1			4	1	4	1
4	1	19	1							4	1	10	1			7	1	21	1
5	1	38	1							5	1	13	1			12	1	27	1
8	1									7	1	19	1			49	1	52	1
78	1									12	2	21	1			59	1		
										13	1	27	1			96	1		
										14	1	35	1						
										41	1	67	1						
										50	1								
										98	1								

The number of nonancestral mutations, n_r, segregating at frequency r in 99 alleles of *Adh* and *Adhr* in *D. pseudoobscura* are shown. Abbreviations for classes of mutations are given in APPENDIX A.

APPENDIX C

Frequency spectra of DNA polymorphism in *D. melanogaster* and *D. simulans*

		D. melanogaster (n = 6)					*D. simulans* (n = 5)				
		n_r					n_r				
Gene	r	Unpref	Pref	MM	NN	Rep	Unpref	Pref	MM	NN	Rep
Adh	1	3	0	0	0	1	4	1	0	0	0
	2	1	0	0	0	0	1	1	0	1	0
	3	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	1	0
	5	3	0	0	0	1					
Adhr	1	0	0	0	0	1	8	2	1	4	0
	2	0	0	0	0	0	1	0	0	1	0
	3	0	0	0	0	0	0	2	0	0	0
	4	0	0	0	0	0	0	1	0	0	0
	5	0	0	0	0	0					
Amy	1	6	1	0	0	2					
	2	4	0	0	1	3					
	3	7	0	0	0	2					
	4	5	0	0	1	0					
	5	0	0	0	0	1					
boss	1	4	0	0	0	2	12	2	0	2	2
	2	1	0	0	1	0	5	0	3	3	0
	3	3	0	0	0	0	2	0	0	1	0
	4	1	0	0	0	0	0	1	0	0	0
	5	2	0	0	2	0					
Mlci	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	1	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0					
per	1	5	0	1	1	1	17	0	1	1	5
	2	0	0	0	2	0	3	0	1	0	1
	3	3	0	0	1	0	0	0	0	0	0
	4	1	0	0	0	0	3	1	0	0	0
	5	5	0	0	0	0					
Pgi	1	1	0	0	0	2	7	3	0	0	2
	2	0	0	0	0	0	2	0	0	0	0
	3	0	0	0	0	0	4	1	0	0	0
	4	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0					
Rh3	1	3	0	2	0	0	7	3	0	2	0
	2	0	0	0	0	0	7	1	0	2	0
	3	0	0	0	0	0	1	0	0	1	0
	4	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0					
Zw	1	4	1	0	0	1	0	1	0	0	0
	2	6	0	0	0	0	3	1	0	0	0
	3	1	0	0	1	0	0	2	0	0	0
	4	0	1	0	0	0	0	0	0	1	0
	5	0	0	0	0	0					

The number of nonancestral mutations, n_r, segregating at frequency r in six alleles of each of nine genes in *D. melanogaster* and five alleles of each of eight genes in *D. simulans* are shown. Abbreviations for classes of mutations are given in APPENDIX A.