The Effects of Mutation and Natural Selection on Codon Bias in the Genes of Drosophila

Richard M. Kliman[1] and Jody Hey

Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received January 21, 1994
Accepted for publication April 23, 1994

ABSTRACT

Codon bias varies widely among the loci of *Drosophila melanogaster*, and some of this diversity has been explained by variation in the strength of natural selection. A study of correlations between intron and coding region base composition shows that variation in mutation pattern also contributes to codon bias variation. This finding is corroborated by an analysis of variance (ANOVA), which shows a tendency for introns from the same gene to be similar in base composition. The strength of base composition correlations between introns and codon third positions is greater for genes with low codon bias than for genes with high codon bias. This pattern can be explained by an overwhelming effect of natural selection, relative to mutation, in highly biased loci. In particular, this correlation is absent when examining fourfold degenerate sites of highly biased genes. In general, it appears that selection acts more strongly in choosing among fourfold degenerate codons than among twofold degenerate codons. Although the results indicate regional variation in mutational bias, no evidence is found for large scale regions of compositional homogeneity.

FOR many organisms, examination of nucleotide sequences of multiple genes has revealed varying levels of the extent to which synonymous codon usage departs from equanimity. In a simple conception, this variation in codon bias is caused by variation in two primary evolutionary forces: (1) mutation, which generates codon diversity, and (2) natural selection against "suboptimal" codons, which reduces codon diversity. The goal of this report is to describe our findings on the relative contributions of mutation and natural selection to the variation in codon bias among the genes of *Drosophila melanogaster*.

In a variety of organisms, studies have shown that variation in codon bias is partly caused by variation among genes in the action of natural selection. In prokaryotes, such as *Escherichia coli*, there is a clear relationship between the extent of codon bias and gene expression level, with more highly expressed genes displaying greater codon bias (GOUY and GAUTIER 1982); a similar pattern has been observed in *Bacillus subtilis* (SHIELDS and SHARP 1987). As would be expected if natural selection limits codon choice, there is also a negative correlation between codon bias and divergence at synonymous sites in *E. coli* and *Salmonella typhimurium* (SHARP and LI 1987). A tendency toward high codon bias in highly expressed genes has also been observed in eukaryotes, such as the yeast *Saccharomyces cerevisiae* (BENNETZEN and HALL 1982) and the slime mold *Dictyostelium discoideum* (SHARP and DEVINE 1989). It appears that natural selection for efficient translation is primarily responsible for biased codon usage, with more highly expressed genes subject to stronger selection pressure than less highly expressed genes (BENNETZEN and HALL 1982; IKEMURA 1985).

For *D. melanogaster*, the evidence for natural selection comes from multiple sources (SHIELDS *et al.* 1988; KLIMAN and HEY 1993). In the first major study on codon bias in this species, SHIELDS *et al.* (1988) presented several pieces of evidence supporting the case that natural selection acts on synonymous codon usage: (1) G + C content is lower in genes with low codon bias, consistent with the expected increase in the influence of the general mutational bias in *D. melanogaster* toward A and T on low biased genes (*i.e.*, selection overcomes mutation pressure in highly biased genes); (2) there is a tendency toward high codon bias in loci homologous to highly expressed (and highly biased) loci in yeast and *E. coli*; (3) among members of a few multigene families, there are anecdotal reports that the more highly biased genes also generate their products in greater abundance; (4) for a small number of cases for which anticodon sequences and tRNA abundances are known, there is a preference in highly biased genes for codons translated by the most abundant iso-accepting tRNA; and (5) the divergence between *D. melanogaster* sequences from homologous *D. pseudoobscura* sequences at synonymous sites is higher in three low biased genes than in three highly biased genes, analogous to the observation in prokaryotes (SHARP and LI 1987). Similarly, in a study that compared 16 homologous sequences from *D. melanogaster* and *D. pseudoobscura*, MORIYAMA and GOJOBORI (1992) found a negative correlation between silent site

[1] Present address: Department of Biology, Radford University, Radford, Virginia 24142.

divergence and codon bias. While there may be natural selection for efficient translation in highly expressed genes in *D. melanogaster*, a recent report that conserved amino acid positions have higher codon bias supports a model in which natural selection also acts on codon usage to maximize the accuracy of translation (AKASHI 1994).

A different tack was taken by KLIMAN and HEY (1993), who used codon bias in *D. melanogaster* to test a population genetics prediction that natural selection is less effective when crossing over is reduced. The central idea is that natural selection could not simultaneously choose from a population the best variants at multiple sites if those sites are tightly linked (HILL and ROBERTSON 1966; FELSENSTEIN 1974; CHARLESWORTH *et al.* 1993). We reasoned that if this prediction is true *and* if natural selection limits codon usage in the highly biased genes of *D. melanogaster*, then codon bias should be lower, on average, in genes located in sections of the genome that experience low levels of crossing over. A large and highly significant difference was found between genes in regions of low recombination and those in the remainder of the genome (KLIMAN and HEY 1993).

Until recently, there has been no evidence from *D. melanogaster* that variation in mutation contributes to variation in codon bias. Several studies have taken the approach of measuring base composition in genomic regions thought to be subject to very low levels of natural selection. In the absence of selection, base composition is a function solely of mutation; however, the function is complex and it is generally not possible to estimate variation in mutation rates among the different nucleotides from information on base composition. SHIELDS *et al.* (1988) compared the G + C content of introns, thought to experience little natural selection on base composition, with the G + C content of silent sites in flanking exons. They did not find a statistically significant correlation [for 36, genes the correlation coefficient was 0.23 ($P = 0.09$)]. Recent studies by MORIYAMA and HARTL (1993) and by CARULLI *et al.* (1993) also found no significant correlation between G + C content at third positions of fourfold degenerate codons and either introns or flanking DNA, respectively. CARULLI *et al.* (1993) did find evidence for compositional heterogeneity across the *D. melanogaster* genome. However, the basis for the heterogeneity could not be established because the proportions of coding sequence and other genetic structures known to influence base composition were not known for the various YAC clones used (CARULLI *et al.* 1993).

The study by KLIMAN and HEY (1993), on the other hand, did show that variation in intron base composition is associated with variation in codon bias. It happens that all of the preferred codons of *D. melanogaster* (*i.e.*, those that appear more frequently in loci with very unequal codon usage) have a G or C in the third position, so that codon bias is highly correlated with G + C content at silent sites (SHIELDS *et al.* 1988; also, see RESULTS). In an analysis of 142 *D. melanogaster* sequences, intron G + C content correlated significantly with codon bias (KLIMAN and HEY 1993). The correlation between intron G + C content and codon bias suggests that the base compositions of introns and silent sites share a common influence, presumably related to regional mutation patterns.

This report further explores the role of mutation, as well as the relative contributions of mutation and natural selection, in the determination of codon usage in *D. melanogaster*.

MATERIALS AND METHODS

Nucleotide sequences: We obtained the complete amino acid-coding sequences of 428 *D. melanogaster* loci from GenBank/EMBL (a table of loci used in this study is available from the authors upon request). Intron sequence was available for 155 of these loci; sequence from more than one intron was available for 79 of the loci.

Codon usage/base composition: Codon usage bias was estimated using the Codon Adaptation Index (CAI; SHARP and LI 1986). This index requires data on codon usage in genes for which natural selection strongly favors a subset of "preferred" codons. We use the codon frequencies presented by SHARP *et al.* (1992) for the most highly biased *D. melanogaster* loci, as determined by their position on the first principal axis produced by correspondence analysis on codon frequencies. For each amino acid, the most common codon used in this subset of genes was assigned a relative frequency of 1.0, and the relative frequency of rarer synonymous codons was scaled accordingly. CAI is calculated as the geometric mean of the relative frequency values corresponding to each codon in a locus. Thus, a gene using only preferred codons would obtain the maximum CAI value of 1.0; genes that use all synonymous codons equally would have a CAI value around 0.2 (the exact value depending on the amino acid composition of the locus); genes with lower values of CAI tend to be biased toward codons that are rare in genes subject to strong selection for optimal codon usage.

The CAI differs in principle from two other commonly used indices of codon bias, Chi/L (SHIELDS *et al.* 1988) and Effective Number of Codons (WRIGHT 1990). The latter indices measure deviation from equal codon usage, regardless of the direction. It should be noted that our choice of CAI rather than the other indices should not substantially affect our results. The three indices are all highly correlated with each other in *D. melanogaster* ($r > 0.92$, $P < 0.0001$ for all pairs of indices using the 428 loci). In general, the extent of deviation from equal codon usage in this species reflects the degree of usage of a particular subset of codons.

In our analyses, fourfold degenerate codons include the fourfold degenerate classes of arginine, leucine and serine. G + C content at third positions of twofold degenerate codons was calculated by summing the number of C-ending codons and the number of G-ending codons (where synonymous choices are either C/T or A/G), and dividing by the total number of twofold degenerate codons. Prior to statistical analyses, all proportion values (*e.g.*, intron G + C content) were arcsine-root transformed.

RESULTS

Codon bias *vs.* G + C content: As previously shown on a smaller sample of loci (SHIELDS *et al.* 1988), G + C

TABLE 1

Base composition correlations

A. Correlations between intron G + C content and G + C content of the three codon positions

	1st position	2nd position	3rd position
2nd position	$r = -0.037$ (NS)		
3rd position	$r = 0.111$ (NS)	-0.099 (NS)	
Intron	$r = 0.245$* ($P = 0.002$)	0.166 ($P = 0.039$)	0.372* ($P < 0.001$)

B. Correlations of intron G + C and individual base content with the base content of the third position of two- and fourfold degenerate codon classes

Base	2-fold	4-fold
A	$r = 0.179$ ($P = 0.026$)	$r = 0.038$ (NS)
C	$r = 0.281$* ($P < 0.001$)	$r = 0.297$* ($P < 0.001$)
G	$r = 0.248$* ($P = 0.002$)	$r = 0.092$ (NS)
T	$r = 0.285$* ($P < 0.001$)	$r = 0.252$* ($P = 0.002$)
G + C	$r = 0.406$* ($P < 0.001$)	$r = 0.269$* ($P = 0.001$)

Statistical significance of Pearson's correlation coefficients before correction for multiple tests is given in parentheses below (NS, $P > 0.1$).

* Significance at $P \leq 0.05$ after correction for six comparisons in A and for 10 comparisons in B, using the sequential Bonferroni method (RICE 1989).

content in third position codon sites is highly correlated to codon bias in *D. melanogaster*. With CAI, Pearson's $r = 0.838$ (426 d.f., $P < 0.0001$); with Chi/L, $r = 0.790$ (426 d.f., $P < 0.0001$); and with Effective Number of Codons, which increases as codon usage becomes more even, $r = -0.798$ (425 d.f., $P < 0.0001$). The slightly weaker correlations involving the latter two indices are due to the small number of loci that are actually biased toward the use of A- and T-ending codons.

Analysis of base composition: If different loci experience different mutation patterns, then positive correlations should be observed between intron base composition and base composition in the coding portions of corresponding exons. This pattern should be most evident in comparisons between introns and codon third positions, where many base substitutions do not cause amino acid changes. From our analyses, intron G + C content correlates significantly with the G + C content in all three codon positions, particularly with that of the third position (Table 1A). As expected, given the latter result, intron G + C content also strongly correlates with CAI ($r = 0.318$, 153 d.f., $P < 0.001$). The simplest interpretation is that base composition is more similar than expected by chance in coding regions of exons and adjacent introns because genomic regions encompassing different loci are subject to different patterns of mutation. An alternative explanation, that natural selection influences the G + C content of both introns and codon third positions in similar ways, can probably be discounted. Loci in areas of reduced recombination have reduced codon bias and reduced codon G + C content, as predicted by models of natural selection (HILL and ROBERTSON 1966; FELSENSTEIN 1974; LI 1987; CHARLESWORTH *et al.* 1993), but the G + C content of the introns does not vary with the level of recombination (KLIMAN and HEY 1993).

The significant G + C correlations conflict with the recent report by MORIYAMA and HARTL (1993) that found no correlation between the base composition of *D. melanogaster* introns and codon third positions. One reason for the different finding is that only fourfold degenerate codons were examined in that study. Table 1B shows that the G + C correlation between introns and third positions of twofold degenerate codons is higher than that involving the fourfold degenerate class, though the two correlation coefficients are not significantly different [$\chi^2 = 1.827$, 1 d.f., $P = 0.176$ (SOKAL and ROHLF 1981, pp. 588–589)]. A second difference between the two studies is sample size. MORIYAMA and HARTL (1993) chose to use only loci with intron sequences of at least 500 bp, in order to avoid biases in intron base composition potentially attributable to conserved signal regions. In contrast, we did not set a lower limit to intron size. However, G + C content of introns, at least in our data set, does not correlate with intron size (Pearson's $r = 0.059$, 153 d.f., $P = 0.464$; Spearman's $r = 0.005$, 153 d.f., $P = 0.951$). Similarly, intron size does not significantly correlate with CAI (Pearson's $r = -0.130$, 153 d.f., $P = 0.107$; Spearman's $r = 0.004$, 153 d.f., $P = 0.960$). It is, therefore, not likely that using short intron sequences has added a systematic bias to any of our analyses on base composition.

Table 1B shows the correlations for each base individually. These correlations are not independent of those involving G + C nor of each other, but, in contrast to the MORIYAMA and HARTL (1993) study on a smaller sample of loci, strong correlations for C and T content were found, even in the fourfold degenerate class. The correlation for G content was also significant, but only for the twofold degenerate class (Table 1B). There is a tendency for stronger correlations with twofold degenerate codons.

Another way to test if different loci are subject to different patterns of mutation is to compare the base composition among introns within genes. We performed an analysis of variance (ANOVA) of base composition within loci, relative to the variance among loci, for the 79 loci with multiple intron sequences. This analysis tests whether the G + C variation among loci with multiple introns is greater than expected given the variation observed between introns within loci. The ANOVA revealed significant variation among loci for intron G + C content (error mean squares = 0.0067; error d.f. = 231; locus level mean squares = 0.0110; locus level d.f. = 78;

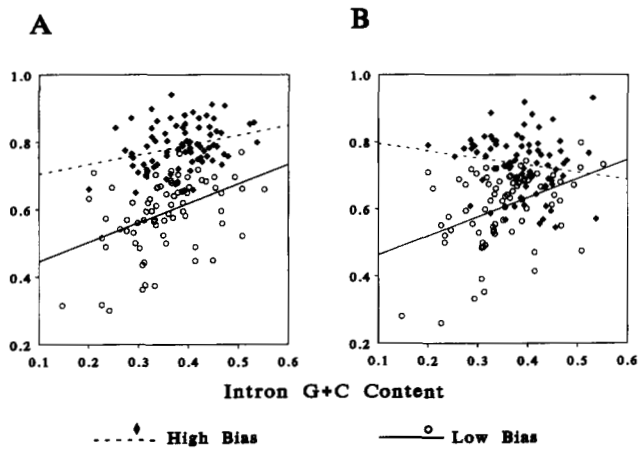

FIGURE 1.—Codon third base G + C content *vs.* intron G + C content. (A) Twofold degenerate codons; (B) fourfold degenerate codons. Linear regression lines are provided as visual aids. Correlation coefficients are given in the text.

$F = 1.644$; $P = 0.0025$), consistent with the results from the correlation analyses.

Effect of selection on mutational bias: The relative contributions of mutation and selection to codon bias variation can be assessed by comparing loci from different portions of the codon bias distribution. The 155 genes were ranked by codon bias, and then split into sets of 77 high and low biased genes (the 78th ranked locus was excluded from this analysis). Spearman's correlation between intron G + C content and codon third position G + C content was calculated for both sets of genes for both twofold and fourfold degenerate codons. The four scatter plots are shown in Figure 1. For the fourfold degenerate codons, the correlation between third base G + C content and intron G + C content is clearly different for the two sets of loci. In the high bias class, Spearman's rank-order correlation is −0.139, while in the low bias class, the correlation is 0.396 (75 d.f., $P < 0.001$); the two correlations differ significantly ($\chi^2 = 11.55$, 1 d.f., $P < 0.00115$). Thus, variation in mutation does not seem to explain G + C variation in fourfold degenerate positions among high bias loci. In contrast, intron G + C content explains about 16% of this variation among low bias genes ($r^2 = 0.157$). For the twofold degenerate class, the correlation for high biased genes is less than that for low biased genes, though the difference is not significant (high bias: Spearman's $r = 0.272$, 75 d.f., $P = 0.017$; low bias: Spearman's $r = 0.399$, 75 d.f., $P < 0.001$; $\chi^2 = 0.76$, 1 d.f., $P = 0.383$). In the low biased genes, the correlations for twofold and fourfold degenerate codons are clearly not different ($r = 0.399$ and $r = 0.396$, respectively); however, in the high biased genes, the correlation for twofold degenerate codons is significantly higher than that for fourfold degenerate codons, even after correcting for multiple tests ($\chi^2 = 6.50$, 1 d.f., $P = 0.011$).

Size of regions of compositional homogeneity: While the previous analyses support a model in which regional mutational biases lead to base composition homogeneity spanning regions at least the size of individual genes, it would be useful to know if larger scale compositional homogeneity exists in the *D. melanogaster* genome. In an effort to uncover potential regions of compositional homogeneity larger than individual loci, we performed a series of MANTEL (1967) tests, first on individual chromosomes, then on the entire data set. The Mantel test is a non-parametric method to test whether two distance (or similarity) matrices are correlated with each other. A correlation coefficient (the normalized Mantel statistic, Z) is calculated and then compared to a distribution of Z values generated by random permutation of one the matrices. In the present application, one matrix is composed of the absolute value of the difference in base composition (either intron G + C or third codon position G + C content) between all pairs of pertinent loci, scaled by division by the largest such value in the matrix. The second matrix contains measures of physical distance between all pairs of loci. This was defined as the amount of DNA between the polytene map positions [as given in the loci.text file of ASHBURNER (1992); if a range of map positions was given, the lower value in the range was used] of two loci divided by the total genomic DNA content, using the estimates of SORSA (1988, pp. 83–105). The frequency with which Z values calculated after permutation equal or exceed the Z value produced by the intact matrices is taken as an estimate of P, the probability that the matrices are similar by chance.

TABLE 2

Mantel tests on G + C content of codon third positions

Chromosome	N	Z	P
2L	75	0.03097	0.2084
2R	70	0.05087	0.1476
3L	73	0.00269	0.4614
3R	123	0.02716	0.1300
X	84	0.11589	0.0018
X^a	61	0.03232	0.1980

N is the number of loci used for the test; also equal to the dimension of the distance matrices. Z is the normalized Mantel statistic calculated from intact data matrices (see text for description). P is the proportion of Z values from permuted matrices that equal or exceed the value from the intact matrices. Five thousand permutations were performed for each test.

[a] Test on X chromosome after removal of loci from polytene bands 1, 2, 19 and 20.

The results of Mantel tests on third position G + C content are given in Table 2. All tests were nonsignificant for both arms of the two major autosomes. A significant result was obtained for the X chromosome. However, a number of loci are found in regions of low recombination, particularly near the tip of the chromosome in polytene bands 1 and 2, and it is known that loci in these regions have reduced codon bias and reduced third codon position G + C content. When loci from polytene bands 1, 2, 19 and 20 were excluded from the analysis (the latter two regions being near the centromere of the X chromosome), the Mantel test was no

TABLE 3

Analysis of heterogeneous base composition within individual loci

Locus	5′ Region[a]	5′ GC:AT	3′ Region[a]	3′ GC:AT	G	P
Gpdh	Introns 1–3	289:374 (0.435)	Introns 4–5	135:254 (0.347)	8.111	0.004
	Exons 2–3	103:25 (0.805)	Exon 5	37:13 (0.740)	0.871	0.351
ptc	Introns 2–3	132:335 (0.283)	Introns 4–5	111:143 (0.437)	17.29	<0.001
	Exon 3	489:153 (0.762)	Exon 5	76:15 (0.835)	2.600	0.107
Su(z)2	Introns 2–3	100:284 (0.260)	Introns 4–5	696:1198 (0.367)	16.73	<0.001
	Exon 3	45:38 (0.542)	Exon 5	675:470 (0.589)	0.709	0.400
trp	Introns 1–6	350:449 (0.438)	Introns 7–13	251:536 (0.319)	23.99	<0.001
	Exons 2–6	281:50 (0.849)	Exons 8–13	336:140 (0.706)	23.10	<0.001

Pooled exon and intron regions for the 5′ and 3′ ends of four genes are provided along with the ratio of G + C to A + T (G + C content in parentheses). For exons, base composition is that of codon third positions only. *G*-tests with one degree of freedom were performed on the G + C *vs.* A + T ratios of the two segments of the gene (separate tests for introns and exons).

[a] Intron and exon boundaries were determined by comparison of nucleotide sequences from genomic DNA and cDNA for each gene [VON KALM *et al.* 1989 (*Gpdh*); HOOPER and SCOTT 1989 (*ptc*); BRUNK and ADLER 1991 (*Su(z)2*); MONTELL and RUBIN 1989 (*trp*)].

longer significant. A test on *X* chromosome intron G + C content was nonsignificant ($Z = 0.05284$, 5000 permutations, $P = 0.2242$), as was a test on all introns ($Z = 0.01971$, 1000 permutations, $P = 0.1290$).

Thus, no evidence for large regions of base composition homogeneity was offered by this set of analyses. It should be noted, however, that our sample of loci is insufficient to detect compositional homogeneity if such regions are smaller than the level of resolution of genome location. We used the lettered chromosome sections provided in the loci.text file of the ASHBURNER (1992) database, and the DNA content of this length of chromosome has an average value of 186 Mbp (SORSA 1988, pp. 83–105). Although the genome is better (*i.e.*, more densely) represented in our samples by codon third positions than by introns, the influence of local mutational bias may be obscured by differences among nearby genes in the strength of natural selection on codon usage, there being no reason to expect, *a priori*, regional similarity in selection (except in the already noted case of regions of low recombination).

Comparison of the base composition among multiple introns from within individual genes does suggest the existence of compositional heterogeneity at a fine scale (*i.e.*, smaller than the size of individual genes). We tested a simplistic model in which G + C variation among introns within genes is caused by each intron having been drawn randomly from the same underlying preference for G + C. *G*-tests of independence on G + C *vs.* A + T content among introns were performed on the 79 loci having more than one intron. The null hypothesis of G + C equanimity among introns was rejected for 33 of the loci (*i.e.*, $P \leq 0.05$; for 17 loci, $P \leq 0.01$). To further examine whether this variation is consistent with a fluctuating mutational pattern within loci, we considered those loci with four or more introns. Of the 18 loci with DNA sequence from at least four introns, four showed a pattern in which at least two adjacent introns with similar G + C content at one end of the gene differed from the remaining introns. This observation might be the result of chance clustering of compositionally similar introns. However, if it is caused by underlying variation in the pattern of mutations, then the G + C content of codon third positions in those exons that lie between introns of similar base composition should also vary, as do the introns, from one end of the gene to the other. We examined third position G + C content in exons "sandwiched" by introns of similar base composition. In all four loci, G + C content in codon third positions was higher in the exons sandwiched by introns with higher G + C content; in the case of the gene encoding *transient receptor potential* (*trp*), the base composition of pooled 5′ exons was significantly greater than that of pooled 3′ exons (Table 3). A simple one-tailed test of whether changing G + C content of third base positions is in the same direction as that of flanking introns is nearly significant (four confirmations out of four comparisons; $0.5^4 = 0.0625$). Thus, it seems that some genes span regions that are heterogeneous in base composition.

DISCUSSION

One way to compare the effects of selection and mutation on codon bias is to estimate the proportions of codon bias variation that can be explained by each of these two causes. By squaring the correlation coefficient between CAI and intron G + C content, we estimate that about 10% ($0.318^2 = 0.10112$) of the variance in codon bias can be explained by the mutational processes that determine base composition. This estimate should be considered as a lower bound, because it assumes that mutation alone, without natural selection, is determining the base composition of introns. In fact, not all nucleotides in introns are expected to be free to mutate, particularly those required for proper intron splicing (GUTHRIE 1991; LESSER and GUTHRIE 1993). Furthermore, recent studies using Drosophila provide evidence for constraint on pre-mRNA secondary structure within intron sequences (STEPHAN and KIRBY 1993; SCHAEFFER and MILLER 1993). However, selection on base

composition in exons (*i.e.*, on synonymous codon usage) is thought to act at the level of translation (Akashi 1994; Bennetzen and Hall 1982; Ikemura 1985), after the introns have been removed. Thus, even if G + C content of introns is partly determined by natural selection, this effect of natural selection is not expected to covary with the selection pressure on adjacent exons. One exception to this is the effect of recombination, whereby loci in regions of reduced crossing over are expected to have reduced selection pressure on both exons and introns. However, the G + C content of introns was not found to vary with the level of recombination (Kliman and Hey 1993).

An estimate of the codon bias variance due to natural selection can be obtained indirectly using recombination, which covaries with the intensity of natural selection. Kliman and Hey (1993) divided loci into two sets based on *a priori* expectations of recombination levels. One set included all loci physically mapped to regions known to experience reduced crossing over, and these loci had markedly reduced levels of codon bias compared to the remaining loci. A single classification ANOVA of codon bias scores grouped by recombination class reveals a highly significant variance due to recombination class (error mean squares = 0.01753, d.f. = 383; recombination class mean squares = 0.25995, d.f. = 1; $F = 14.83$; $P < 0.001$). Following Sokal and Rohlf (1982, p. 216), 16% of the variance in codon bias is explained by recombination class. This quantity is also a lower bound (probably even more so than the estimate of the variance due to mutation), because of the very indirect method of equating recombination classes with the intensity of natural selection.

Two of our analyses suggest that natural selection on the twofold degenerate class of codons is weak relative to selection on the fourfold degenerate class. First, base composition of the twofold degenerate class tends to correlate more strongly with that of introns than does that of the fourfold degenerate class (Table 1B). Second, the third codon position G + C content of the twofold degenerate class correlates significantly with that of introns even in the set of genes with high codon bias. In contrast, no G + C correlation was found between introns and third codon positions of fourfold degenerate codons in highly biased genes, consistent with an overriding effect of natural selection on fourfold degenerate codon usage. The difference between two- and fourfold degenerate codons, in addition to sample size considerations, probably explains why two previous studies that focused only on fourfold degenerate codons failed to support a model of regional variation in mutation patterns (Moriyama and Hartl 1993; Carulli *et al.* 1993).

An explanation for the difference in natural selection between twofold and fourfold degenerate codon classes comes from information on tRNA levels in other organisms. In *E. coli*, *S. typhimurium* and *S. cerevisiae*, twofold degenerate codons usually use a single tRNA, while those with higher degeneracy usually use more than one tRNA (Ikemura 1985). If the tRNA repertoire of *D. melanogaster* is similar, then natural selection may have more scope within those codon classes that employ multiple tRNAs. In the case of the fourfold degenerate class, selection may favor the codon (or codons) that uses the most common iso-accepting tRNA and/or bind that tRNA best. However, in the case of the twofold degenerate class, there may be less, if any, selection for iso-accepting tRNA usage. Therefore, selection coefficients for optimal codon usage in the twofold degenerate class of amino acids may be lower than those for the fourfold degenerate class. It should be noted, however, that twofold degenerate codons are clearly influenced by natural selection to some extent. G + C content of third positions of twofold degenerate codons correlates much more strongly with that of fourfold degenerate codons (Pearson's $r = 0.664$, 153 d.f., $P < 0.001$) than can possibly be explained by the apparently weak effect of regional mutational bias, indicating that genes that are highly biased for fourfold degenerate codons are also highly biased for twofold degenerate codons.

An alternative reason for the higher correlations in low bias genes is that some of the introns of these genes contain unidentified amino acid coding regions and have, therefore, been subject to selection pressure similar to that of flanking exons. Depending on how the intron boundaries were characterized in the original reports, this could come about if there are multiple splicing protocols for a locus (though an effort was made to avoid such loci in this study) or if intron position was mistakenly inferred from the genomic DNA sequence. A third possibility, that some introns contain genes on the opposite strand, can probably be ruled out given the short length of Drosophila introns (the median length of intron sequences used in this study is 66 bp; the mean length was 271 bp; of the 389 introns examined, 52 exceeded 500 bp and 21 exceeded 1000 bp). However, comparisons of absolute G + C contents between introns and flanking exons make the presence of unidentified coding sequence unlikely. In all 155 loci studied, G + C content of coding regions exceeds that of associated introns. More specifically, in 154 cases, G + C content of codon third positions (*i.e.*, considering all such positions, only twofold degenerate sites or only fourfold degenerate sites) is greater than that of associated introns. These observations are expected if silent sites are under selection [an argument made before by Shields *et al.* (1993)]; they also lead to the prediction that the presence of unidentified coding sequence in introns should lead to an increase in average intron G + C content. If the introns of low biased genes, in particular, contain unidentified coding sequence, we would expect, on average, greater intron G + C content

in low biased genes. However, this is not observed. Average intron G + C content in the 77 genes with lower codon bias is 0.351, while that of the 77 genes with higher codon bias is 0.386.

The cause of base composition heterogeneity among *D. melanogaster* loci is not clear. There is evidence for heterogeneity in average G + C content of large *D. melanogaster* genome segments cloned into yeast artificial chromosomes (YAC) (CARULLI *et al.* 1993), but no pattern analogous to the isochores of warm-blooded vertebrates [*i.e.*, base composition homogeneity over megabase stretches of the genome (BERNARDI and BERNARDI 1985; BERNARDI 1989)] has yet been found in this species. That is, the study by CARULLI *et al.* (1993) did not indicate that heterogeneity in the composition of YAC clones was attributable to differences in mutational bias or large scale selection on base composition. Some hypotheses proposed for vertebrate isochore formation can probably be ruled out. The theory that isochores result from genome level natural selection related to homeothermy is plainly inappropriate for *D. melanogaster* (BERNARDI *et al.* 1985; BERNARDI and BERNARDI 1986). A relationship between recombination rate and base composition observed in mammals, perhaps involving biased gene conversion (EYRE-WALKER 1993), is also unlikely. Our previous report found no difference in intron G + C content when regions of low recombination were compared to other regions of the genome (KLIMAN and HEY).

Several theories of isochore formation invoke a role for variation in the timing of DNA replication, though the actual factor influencing base composition is unclear (HOLMQUIST 1987, 1989; WOLFE *et al.* 1989; WOLFE 1991). DNA replication appears to begin synchronously in adjacent chromosomal regions in vertebrates (HOLMQUIST 1987), which, when coupled with a replication time effect, could explain why long stretches of the genome have similar base composition. In contrast, adjacent replication forks do not arise synchronously in Drosophila (STEINEMANN 1981). Therefore, it remains possible, in principle, that large scale vertebrate isochores and local mutational biases in *D. melanogaster* arise through similar mechanisms, with the major difference being the size of synchronously replicating genomic regions.

We thank A. C. EYRE-WALKER, A. BARBADILLA, R. C. LEWONTIN, E. N. MORIYAMA and D. L. HARTL for helpful discussion. We also thank H. AKASHI for valuable comments on the manuscript. This research was supported by grant BSR8918164 to J.H. from the National Science Foundation.

LITERATURE CITED

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

ASHBURNER, M., 1992 Flybase: A Drosophila Genetic Database, Version 9209 (available electronically from the ftp.bio.indiana.edu and EMBL file servers).

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

BERNARDI, G. 1989 The isochore organization of the human genome. Annu. Rev. Genet. **23:** 637–661.

BERNARDI, G., and G. BERNARDI, 1985 Codon usage and genome composition. J. Mol. Evol. **22:** 363–365.

BERNARDI, G., and G. BERNARDI, 1986 Compositional constraints and genome evolution. J. Mol. Evol. **24:** 1–11.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

BRUNK, B. P., and P. N. ADLER, 1991 The sequence of the Drosophila regulatory gene *Suppressor two of zeste*. Nucleic Acids Res. **19:** 3149.

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., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular evolution. Genetics **134:** 1289–1303.

EYRE-WALKER, A., 1993 Recombination and mammalian genome evolution. Proc. R. Soc. Lond. Ser. B **252:** 237–243.

FELSENSTEIN, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GUTHRIE, C., 1991 Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science **253:** 157–163.

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

HOLMQUIST, G. P., 1987 Role of replication time in the control of tissue-specific gene expression. Am. J. Hum. Genet. **40:** 151–173.

HOLMQUIST, G. P., 1989 Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. **28:** 469–486.

HOOPER, J. E., and M. P. SCOTT, 1989 The Drosophila *patched* gene encodes a putative membrane protein required for segmental patterning. Cell **59:** 751–765.

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

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

LESSER, C. F., and C. GUTHRIE, 1993 Mutations in U6 snRNA that alter splice site specificity: implications for the active site. Science **262:** 1982–1988.

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

MANTEL, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

MONTELL, C., and G. M. RUBIN, 1989 Molecular characterization of the Drosophila *trp* locus: a putative integral membrane protein required for phototransduction. Neuron **2:** 1313–1323.

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

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

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

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 K. M. DEVINE, 1989 Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do "prefer" optimal codons. Nucleic Acids Res. **17:** 5029–5038.

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., C. J. BURGESS, A. T. LLOYD and K. J. MITCHELL, 1992 Selective use of termination codons and variations in codon choice, pp. 397–425 in *Transfer RNA in Protein Synthesis*, edited

by D. L. HATFIELD, B. J. LEE and R. M. PIRTLE. CRC Press, Boca Raton, Fla.

SHIELDS, D. C., and P. M. SHARP, 1987 Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. **15:** 8023–8040.

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 & Co., New York.

SORSA, V., 1988 *Chromosome Maps of Drosophila*, Vol. 2. CRC Press, Boca Raton, Fla.

STEINEMANN, M., 1981 Chromosomal replication in *Drosophila virilis*. III. Organization of active origins in the highly polytene salivary gland cells. Chromosoma **82:** 289–307.

STEPHAN, W., and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135:** 97–103.

VON KALM, L., J. WEAVER, J. DEMARCO, R. J. MACINTYRE and D. T. SULLIVAN, 1989 Structural characterization of the α-glycerol-3-phosphate dehydrogenase-encoding gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 5020–5024.

WOLFE, K. H., 1991 Mammalian DNA replication: mutation biases and the mutation rate. J. Theor. Biol. **149:** 441–451.

WOLFE, K. H., P. M. SHARP and W.-H. LI, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

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

Communicating editor: D. CHARLESWORTH