Codon Usage Bias and Base Composition of Nuclear Genes in Drosophila

Etsuko N. Moriyama[1] and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138

Manuscript received October 28, 1992
Accepted for publication March 25, 1993

ABSTRACT

The nuclear genes of Drosophila evolve at various rates. This variation seems to correlate with codon-usage bias. In order to elucidate the determining factors of the various evolutionary rates and codon-usage bias in the Drosophila nuclear genome, we compared patterns of codon-usage bias with base compositions of exons and introns. Our results clearly show the existence of selective constraints at the translational level for synonymous (silent) sites and, on the other hand, the neutrality or near neutrality of long stretches of nucleotide sequence within noncoding regions. These features were found for comparisons among nuclear genes in a particular species (*Drosophila melanogaster*, *Drosophila pseudoobscura* and *Drosophila virilis*) as well as in a particular gene (alcohol dehydrogenase) among different species in the genus *Drosophila*. The patterns of evolution of synonymous sites in Drosophila are more similar to those in the prokaryotes than they are to those in mammals. If a difference in the level of expression of each gene is a main reason for the difference in the degree of selective constraint, the evolution of synonymous sites of Drosophila genes would be sensitive to the level of expression among genes and would change as the level of expression becomes altered in different species. Our analysis verifies these predictions and also identifies additional selective constraints at the translational level in Drosophila.

SYNONYMOUS sites of coding sequences have been suggested to be nearly neutral for base substitutions (KIMURA 1983) and to evolve at almost the same rate among different genes and also among different organisms (MIYATA, YASUNAGA and NISHIDA 1980) because base substitutions at synonymous sites do not change amino acid sequences. However, as more sequence data have become available from molecular studies, there has been shown to be considerable variation in the rates of synonymous substitution among taxonomic lineages (BRITTEN 1986; MORIYAMA 1987) and even among genes in the same genome (SHARP and LI 1989). For variation among lineages, there may be an effect of difference in generation time. On the other hand, variation within the same genome might be explained by differences in selective constraints and/or mutation rates. Recent observations that synonymous codons are not used at equal frequency and that codon usage varies among organisms and even among genes in the same organism (IKEMURA 1985) suggest that the synonymous sites in DNA sequences could be subject to some selective constraints at the translational level.

In prokaryotes, a negative correlation between the rates of synonymous substitution and the codon-usage bias has been shown for the comparison between *Escherichia coli* and *Salmonella typhimurium* (SHARP and LI 1987; SHARP 1991). Very highly expressed genes have high codon-usage bias and have diverged very little; very lowly expressed genes have little codon-usage bias and high synonymous substitution rates. These observations clearly show that synonymous sites in these bacterial genes are subject to different degrees of selective constraint at the translational level, such as tRNA abundance. In mammals, on the other hand, the substitution rates at synonymous sites seem to have a relationship with the G + C content in each chromosomal region. Comparisons between two species of rodents (mouse and rat) have indicated that genes with intermediate G + C content at synonymous sites are most divergent. Genes with 50% or higher G + C content have a strong negative correlation between rates of synonymous substitution and G + C content, whereas genes with G + C content lower than 50% have a positive correlation. Such variation in synonymous substitution rates among mammalian genes seem to reflect differences in the rates and patterns of mutation over the mammalian genome rather than differences of selective constraints on synonymous codons (BULMER 1987; WOLFE, SHARP and LI 1989).

Synonymous sites in nuclear genes of Drosophila also evolve at highly variable rates. Such variation has been shown not only for the comparison among different genes of the same Drosophila species, but also for the comparison of a particular gene (*e.g.*, alcohol dehydrogenase) among different Drosophila lineages (SHIELDS *et al.* 1988; SHARP and LI 1989; MORIYAMA

[1] On leave from: Department of Evolutionary Genetics, National Institute of Genetics, Mishima 411, Japan.

and GOJOBORI 1992). The synonymous substitution rates of Drosophila nuclear genes have a negative correlation with codon-usage bias. Whether or not such a relationship can be explained by differences in selective constraints at the translational level, as described for enterobacterial genes, is a controversial issue at present, because Drosophila is a multicellular eukaryote, and the expression level of each gene could be different among tissues and/or among developmental stages.

In a previous paper (MORIYAMA and GOJOBORI 1992), we showed that there are strong correlations between synonymous substitution rates and base compositions at synonymous sites. We presented several alternative explanations for the result. However, we could not determine whether the most plausible explanation was related to differences in selective constraints on synonymous codons at the translational level, differences in selective constraints on some types of nucleotide changes, or differences in underlying mutation rates or patterns of mutation. In the absence of sufficient data about tRNA abundance and level of gene expression in Drosophila, we could not determine definitively whether differences in selective constraints at the translational level cause the codon-usage bias in Drosophila. We compared the base composition of pseudogenes with that of the functional homologs to examine the neutral state. However, the number of pseudogenes was limited and, moreover, we could not guarantee the neutrality of pseudogenes because we do not know their early evolutionary history.

To identify the main cause of the codon-usage bias in Drosophila, we should compare nucleotide sequences under a functional constraint with those under no or very weak constraint, under conditions in which both types of sequences have the same rates and pattern of mutation. In this study, we use long introns to provide a better model for a neutral nucleotide sequence than pseudogenes used previously. In particular, if differences in mutation rates or patterns of mutation in different regions of the Drosophila nuclear genome cause the variation in base composition and synonymous substitution rates, the base composition of introns of a gene should be almost the same as that of synonymous sites in exons of the same gene. We have therefore examined the relationship between the base composition of introns with that of synonymous sites in adjacent exons for a number of Drosophila nuclear genes, and we have carried out detailed analysis of codon-usage bias. The results are discussed in light of the mechanisms underlying codon-usage bias, base composition, and base substitutions at synonymous sites.

MATERIALS AND METHODS

Nucleotide sequences of introns: Nucleotide sequences of introns contain several signals for splicing or other regulation. These signal regions are relatively conserved in nucleotide sequences and may produce a bias in any analysis of base composition that includes these sequences. To minimize such bias, we used only genes with introns greater than 500 base pairs (bp). For *Drosophila melanogaster*, we examined 18 genes with introns of 500–1000 bp and 22 genes with introns greater than 1000 bp obtained from GenBank DNA database 71.0 (see APPENDIX, Table 3).

Nucleotide sequences of synonymous sites: To avoid biases in base compositions resulting from biases in amino acid composition, we used only the nucleotide sequences at fourfold degenerate sites for calculating base compositions of synonymous sites. (A fourfold degenerate site allows any nucleotide without changing the amino acid.) Furthermore, to determine whether the pattern of base composition or codon usage differs among synonymous codon groups, we also examined the relation between the base composition of fourfold degenerate sites and codon-usage bias for each synonymous codon group. Synonymous codon groups with less than 10 codons in each gene were excluded from this analysis.

Chromosomal location of each gene: Spatial patterns of base compositions and codon-usage bias over the nuclear genome of Drosophila were examined. Cytological map positions of the genes were taken from ASHBURNER and GELBART (1991, see APPENDIX, Table 3).

Nucleotide sequences of alcohol dehydrogenase genes: We compared base compositions of alcohol dehydrogenase (*Adh*) genes among total of 34 Drosophila species: eight from the *melanogaster* group, five from the *obscura* group, one from the *willistoni* group, nine from the *repleta* group, nine from the Hawaiian *Drosophila*, one from the Hawaiian species of the subgenus *Engiscaptodrosophila*, and one from the subgenus *Scaptodrosophila* (see Table 2). The nucleotide sequences were obtained from GenBank DNA database 71.0 (APPENDIX, Table 4), except for *D. ambigua* (MARFANY and GONZALEZ-DUARTE 1991) and *D. willistoni* (ANDERSON, CAREW and POWELL 1993). We analyzed the longest intron (500–900 bp) in the 5′ region of the *Adh* genes for species of the *melanogaster* and *obscura* groups and for *D. affinidisjuncta*; the flanking noncoding sequences (1100–3200 bp) between *Adh-2* and *Adh-1* for species of the *repleta* group; and the 3′ flanking sequences (about 380 bp) for species of the Hawaiian picture-winged group. For the *repleta* group, we used only the coding sequences of the *Adh-1* gene, although these species have duplicated *Adh* genes (*Adh-1* and *Adh-2*).

Codon-usage bias: We used the method of SHIELDS *et al.* (1988) for calculating the degree of codon-usage bias. Their value of "scaled χ^2" is calculated as the χ^2 value, computed for the deviation from equal usage of codons within a synonymous codon group, divided by the total number of codons in the gene (L), excluding Trp, Met, and termination codons.

RESULTS

Relations between codon-usage bias and base composition: The "scaled χ^2" (χ^2/L, where L is the number of codons) is a measure of codon-usage bias (SHIELDS *et al.* 1988). Scaled χ^2 has a negative correlation with the rate of synonymous substitution. However, the scaled χ^2 indicates only the degree of bias. It does not indicate the nature or pattern of codon-usage bias. We have examined the relations between base composition and synonymous substitution rates in a previous paper (MORIYAMA and GOJOBORI 1992).

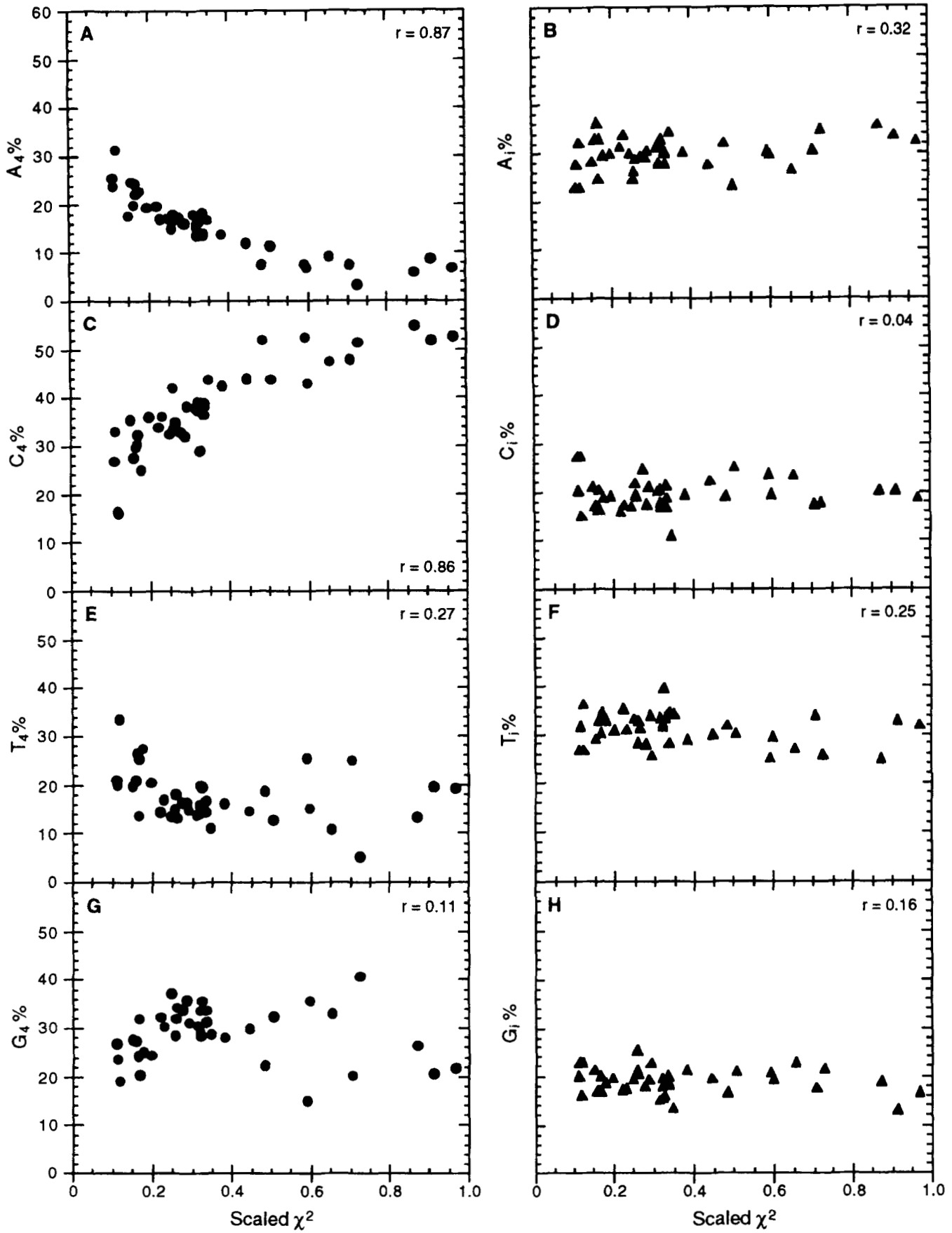

FIGURE 1.—Relation between scaled χ^2 and the identity of the third nucleotide at fourfold degenerate sites (A, C, E and G) and base composition of introns (B, D, F and H) for nuclear genes of *D. melanogaster*. Forty genes with introns greater than 500 bp were used for calculation (see MATERIALS AND METHODS). $A_4\%$, $C_4\%$, $T_4\%$ and $G_4\%$ are the percentage of fourfold degenerate codons that end in A, C, T or G, respectively; $A_i\%$, $C_i\%$, $T_i\%$ and $G_i\%$ are the percentage of A, C, T, and G, respectively, in introns longer than 500 bp.

TABLE 1

Relation between codon-usage bias and base content for each fourfold degenerate codon group

Codon group (amino acid)	No. of genes[a]	Regression coefficient[b]	
		NNA	NNC
CUN (Leu)	59	−0.65**	0.06[ns]
GUN (Val)	61	−0.50**	0.48**
UCN (Ser)	57	−0.72**	0.61**
CCN (Pro)	56	−0.35*	0.78**
ACN (Thr)	62	−0.55**	0.77**
GCN (Ala)	64	−0.59**	0.76**
CGN (Arg)	56	−0.63**	0.46**
GGN (Gly)	64	−0.50**	0.42**

Scaled χ^2 was calculated for the coding sequence of each gene.
[a] Number of genes having more than 10 codons for each fourfold degenerate codon group.
[b] * Significant at 1% level; ** significant at 0.1% level; [ns] not significant.

In this paper, we first present the details of codon-usage bias in the Drosophila nuclear genes. Figure 1 (A, C, E and G) shows the relations between the base compositions of fourfold degenerate sites and scaled χ^2 for genes in *D. melanogaster*. The symbols A_4%, C_4%, T_4% and G_4% represent the percentage of fourfold degenerate codons that end in A, C, T or G, respectively. The degree of codon-usage bias was correlated strongly positive with the C content of fourfold degenerate sites (C_4%, $r = 0.86$, $P < 0.001$; Figure 1A) and negatively correlated with the A_4% ($r = -0.87$, $P < 0.001$; Figure 1C). We should note that all of these figures include a significant number of highly biased genes. Excluding the highly biased genes (scaled $\chi^2 > 0.5$), the correlation coefficients (r) for T_4 and G_4% were −0.56 ($P < 0.01$) and 0.35 ($P > 0.05$), respectively. Therefore, we can say that high codon-usage bias of Drosophila nuclear genes is associated with high C and low A contents at synonymous sites. There is also a weaker tendency toward avoidance of T.

Patterns of codon-usage bias in each synonymous codon group: In order to clarify whether there is an effect of the amino acid composition on the bias of codon usage or base composition, we examined the relations between the scaled χ^2 and base composition at the third nucleotide of each synonymous codon group (Table 1 and Figure 2). For all eight of the fourfold degenerate synonymous codon groups, we found the same relations between A and C content and scaled χ^2 as shown in Figure 1, although a few synonymous codon groups showed no or weak correlation (the groups CUN:Leu and CCN:Pro in Table 1). Therefore, the relations between base composition and scaled χ^2 are not a specific feature for a particular amino acid or synonymous codon group, but reflect a more general phenomenon.

Distribution of codon-usage bias within genes: We divided the coding region of each gene into three almost equal segments and calculated the base composition and scaled χ^2 for each gene segment. For bacterial genes, a lower codon-usage bias has been shown at the 5′ ends of the coding regions (BULMER 1988; LAWRENCE and HARTL 1991). Such a spatial bias of codon usage may effect the rates of translation of the genes (LILJENSTRÖM and VON HEIJNE 1987). For Drosophila nuclear genes, however, unlike enterobacterial genes, there was no particular spatial pattern of codon usage (data not shown).

Base composition of introns: Figure 1, D, F and H, indicates that the base contents of introns (denoted C_i, T_i and G_i% in Figure 1) have no correlation with either the base content at synonymous sites (C_4, T_4 and G_4%) or the scaled χ^2. However, a slight positive correlation was shown for A_i% (Figure 1B), although this correlation disappeared when we examined only introns greater than 1000 bp. Moreover, the introns were essentially constant in base composition and virtually independent of the base composition at synonymous coding sites in the same genes. The average base composition of the nontranscribed strand of introns of *D. melanogaster* is 30.1% A, 31.4% T, 19.4% C and 18.9% G. The base composition of introns is almost the same as expected if codon usage at synonymous sites were not biased (*i.e.*, the extrapolated base composition obtained when the scaled χ^2 equals 0). Judging from this aspect of the base composition, the introns of Drosophila nuclear genes are not subject to selective constraints of the same sort that result in the codon-usage bias of synonymous sites. The long (>500 bp) introns of Drosophila appear to evolve in a nearly neutral fashion.

Constancy of base composition over the Drosophila nuclear genome: We also examined the spatial pattern of base composition over the Drosophila nuclear genome. There was no detectable relation between the scaled χ^2 and chromosomal location or between the base composition of fourfold degenerate sites and chromosomal location (data not shown). Therefore, the codon-usage bias shown here is not caused by bias in selective constraints or mutation rates as a function of chromosomal location. This finding supports other evidence against an isochore structure or other higher-order compartmentalization of different classes of sequence in the Drosophila genome (CARULLI *et al.* 1993).

Base compositions of the other Drosophila species: Similar relations as those found in *D. melanogaster* were also shown for both *D. pseudoobscura* and *D. virilis* (Figure 3). The base compositions of introns have quite constant values among genes, whereas the base compositions of fourfold degenerate sites show a correlation with the degree of codon-usage bias. The average base contents of introns of both *D. pseudoobscura* and *D. virilis* were identical to those of *D. melanogaster*. Therefore, the base compositions of in-

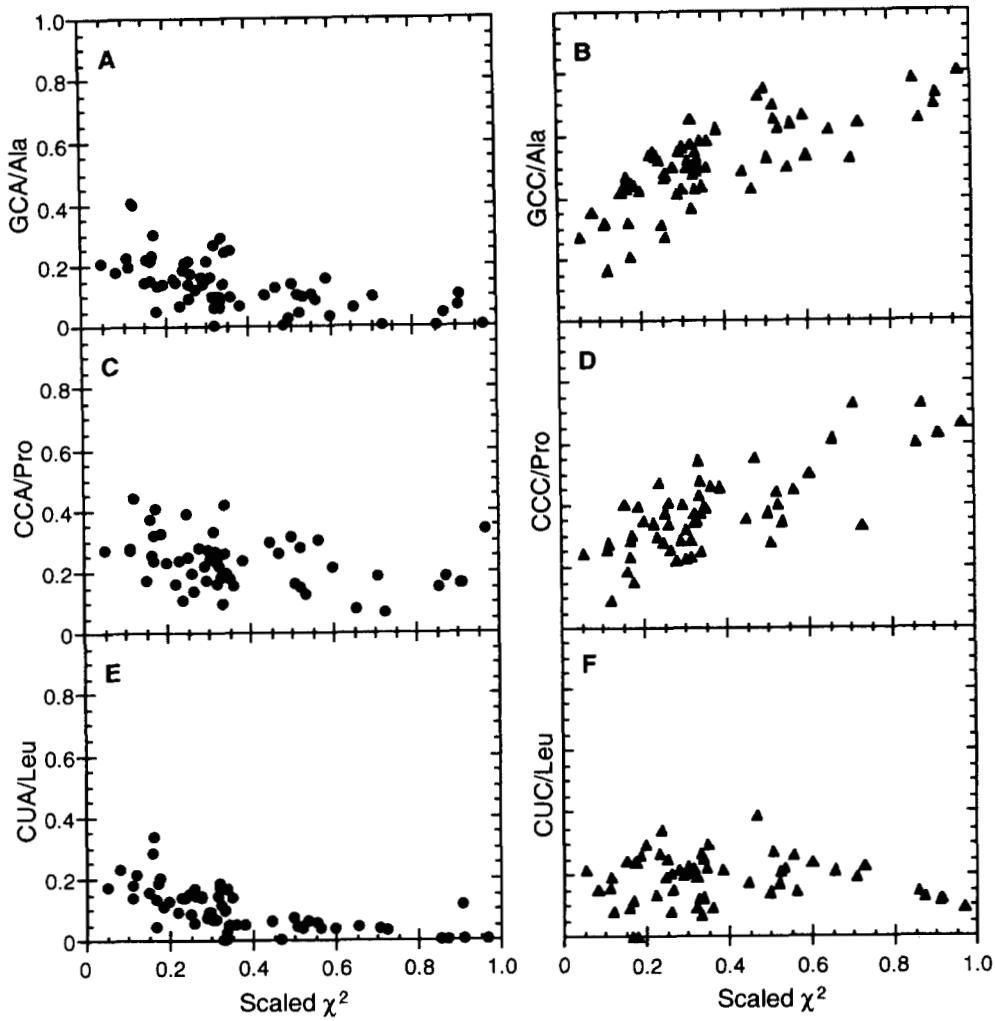

FIGURE 2.—Relation between scaled χ^2 and the identity of the third nucleotide in fourfold degenerate codon groups (GCN, A and B; CCN, C and D; CUN, E and F). For regression coefficients and other synonymous groups, see Table 1.

trons seem to be constant over the Drosophila nuclear genome and also over these widely divergent Drosophila lineages.

Constancy of base composition over Drosophila lineages: In order to examine the constancy of base composition over a more extensive sample of Drosophila lineages, we examined the base composition of the *Adh* genes among 34 Drosophila species. We obtained results similar to those noted earlier (Table 2 and Figure 4). The codon usage and base composition of fourfold degenerate sites of *Adh* genes were widely different among Drosophila species, as mentioned also in a previous paper (MORIYAMA and GOJOBORI 1992). However, the base compositions of introns and noncoding regions of the *Adh* genes were remarkably constant over the *Sophophora*, *Drosophila*, *Engiscaptodrosophila* and *Scaptodrosophila*. Therefore, it is possible that the base composition is constant over the nuclear genomes of all Drosophila lineages and may be maintained by very general patterns of mutation or DNA repair.

DISCUSSION

In mammalian genomes, both the mosaic structure of the genome (isochores) and variation in synonymous substitution rates seem to be caused by regional differences in the pattern of mutation (WOLFE, SHARP and LI 1989) rather than by selective constraints at the translational level. The strong correlation between the G + C contents of synonymous sites and those of introns and flanking noncoding regions (AOTA and IKEMURA 1986; BULMER 1987) supports this model. On the other hand, in enterobacteria, a clear correlation between codon-usage bias and synonymous substitution rate has been shown. This correlation could be explained by selective constraints on synonymous codons rather than regional differences in mutation rates. Although selection coefficients for synonymous codons are expected to be very small, the effective population size of bacteria is so large that the slight difference of selection coefficients could produce a bias in codon usage (LI 1987; SHARP 1989). As expected, the base composition of 5′ or 3′ flanking regions is not correlated with that of synonymous sites in enterobacterial genes (SHARP 1990).

Drosophila does not seem to have a large-scale isochore structure comparable to that observed in mammals, although some compartmentalization was shown among clones the size of yeast artificial chromosomes (~200 kb) (CARULLI *et al.* 1993). If there

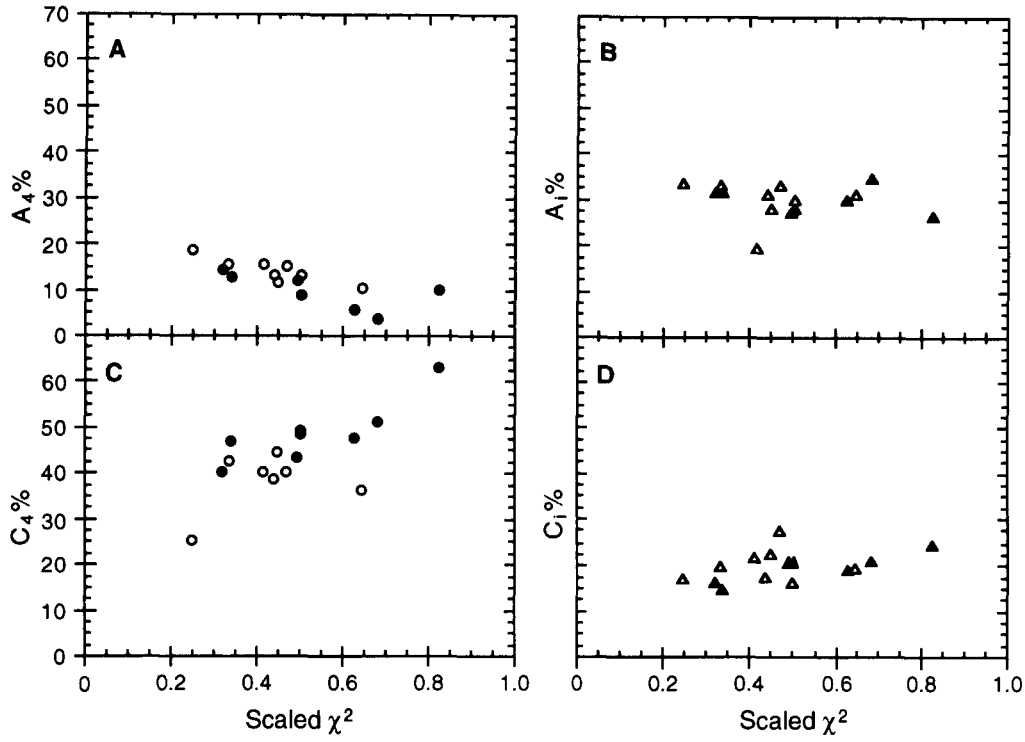

FIGURE 3.—Relation between scaled χ^2 and the identity of the third nucleotide in fourfold degenerate codons (A and C) and base composition of introns (B and D) for nuclear genes of *D. pseudoobscura* (● and ▲) and *D. virilis* (○ and △). A_4%, C_4%, A_i% and C_i% are defined as in Figure 1.

were smaller regions with biased base composition in Drosophila resulting from biased mutation patterns, then one would expect to find homogeneous base compositions surrounding genes and heterogeneous base compositions in different regions. However, the results obtained here show no correlation between base composition of introns and flanking noncoding regions with those at fourfold degenerate sites in coding regions. The base composition of introns was surprisingly constant over the Drosophila nuclear genome and also over Drosophila lineages. Under the assumption that the rates and patterns of mutation in introns are the same as those of synonymous sites in the coding regions, we conclude that the variation in synonymous substitution rates among Drosophila nuclear genes reflects differences in selective constraints for synonymous codons at the translational level rather than any spatial bias in mutation pattern. On the other hand, the nucleotide sequences of introns in Drosophila genes seem to be virtually random in base composition (excluding short sequences required for splicing, processing, enhancers, or other functions). Therefore, the underlying pattern of mutation in the Drosophila nuclear genome appears to be very similar over the genome and over the genus. In this respect, the evolutionary features of the Drosophila nuclear genome seem to be similar to those of bacterial genes.

With respect to codon-usage bias, SHIELDS *et al.* (1988) suggested that the effective population size of Drosophila species may be sufficiently large (10^6–10^7) that differences in selection coefficients of synonymous codons are significant. However, there are other issues related to codon-usage bias. For example, the synonymous substitution rates in Drosophila are correlated with the scaled χ^2, which is the value representing the degree of deviation from equal usage of synonymous codons. However, the scaled χ^2 is different from the CAI (codon adaptation index) used for bacterial genes, which is a measure of codon-usage bias relative to a reference (postulated to be optimal) codon usage (SHARP and LI 1987). The correlation of synonymous substitution rates with scaled χ^2 suggests that the preferable (maybe optimal) codon usage in Drosophila is expected to be virtually the same for all nuclear genes. However, both the expression level of genes and tRNA abundance must vary with tissue and/or developmental stage, and thus the preferred synonymous codons might also be expected to vary with tissue, developmental stage, and gene. Therefore, it is remarkable that synonymous substitution rates of Drosophila nuclear genes have a simple correlation with codon-usage bias. On the other hand, the codon-usage bias described here reflects an average across genes subjected to various selective constraints in various tissues at various developmental stages, and so the apparent homogeneity may be misleading.

The average base composition of the nontranscribed strand of introns was 30.1% A, 31.4% T,

TABLE 2

Scaled χ^2 and base compositions at synonymous sites and introns of *Adh* genes of Drosophila species

Species	Scaled χ^2	Synonymous sites				Introns			
		A_4%	T_4%	C_4%	G_4%	A_i%	T_i%	C_i%	G_i%
Subgenus *Sophophora*									
(*melanogaster* group)									
D. melanogaster	0.87	5.8	13.1	54.7	26.3	35.8	24.9	20.3	19.0
D. mauritiana	0.94	5.8	13.8	55.1	25.4	36.1	24.4	20.8	18.8
D. sechellia	0.92	5.1	15.2	54.4	25.4	36.3	24.7	19.9	19.1
D. simulans	0.92	5.1	13.8	54.4	26.8	36.4	24.2	20.6	18.8
D. erecta	0.93	5.1	10.1	56.5	28.3	36.6	21.5	21.3	20.5
D. orena	0.88	5.2	11.8	54.4	28.7	36.4	21.5	20.4	21.8
D. yakuba	1.00	5.1	13.1	51.8	29.9	35.9	21.8	21.3	21.1
D. teissieri	1.07	5.1	9.4	55.8	29.7	35.6	23.7	21.2	19.6
(*obscura* group)									
D. pseudoobscura	0.68	3.7	18.3	51.1	27.0	34.9	25.7	21.2	18.3
D. persimilis	0.67	3.7	17.7	52.2	26.5	35.1	25.5	21.3	18.1
D. miranda	0.65	3.7	19.1	51.5	25.7	35.6	25.7	20.9	17.8
D. subobscura	0.46	11.2	21.6	41.0	26.1	33.3	25.1	23.0	18.6
D. ambigua	0.94	3.0	11.9	50.4	34.8	33.1	23.3	22.5	21.0
(*willistoni* group)									
D. willistoni	0.45	11.5	30.3	43.4	14.8				
Subgenus *Drosophila*									
(*repleta* group)									
D. mulleri	0.51	12.8	17.3	44.4	25.6	33.8	32.4	16.0	17.9
D. arizonae[a]	0.42	16.4	19.4	39.6	24.6				
D. mayaguana[a]	0.42	15.8	22.6	36.1	25.6				
D. buzzatii[a]	0.29	18.5	23.7	36.3	21.5				
D. wheeleri[a]	0.34	14.4	20.5	39.4	25.8				
D. mojavensis	0.34	14.6	18.5	42.3	24.6	32.8	31.8	16.6	18.9
D. mettleri	0.36	15.2	23.5	36.4	25.0	36.4	30.6	16.4	16.7
D. hydei	0.32	16.5	26.3	36.8	20.3	36.0	31.2	15.3	17.5
D. navojoa	0.42	15.3	14.5	46.6	23.7	31.6	30.1	17.4	20.8
(Hawaiian picture-winged group)									
D. adiastola	0.31	15.0	29.3	36.1	19.6				
D. picticornis	0.33	16.9	23.9	37.7	21.5	36.0	29.0	16.8	18.1
D. affinidisjuncta	0.31	17.9	27.6	32.8	21.6	40.5	25.1	17.3	17.1
D. planitibia	0.34	14.9	26.1	36.6	22.4	38.5	28.3	15.5	17.8
D. heteroneura	0.33	15.2	26.5	37.1	21.2	37.5	27.8	15.8	18.9
D. silvestris	0.33	15.3	26.0	37.4	21.4	36.7	28.3	15.5	19.6
D. differens	0.33	15.7	26.1	35.8	22.4	38.6	28.2	15.1	18.0
(Hawaiian modified-mouthparts group)									
D. mimica	0.33	11.3	29.3	35.3	24.1				
(Hawaiian fungus-feeders group)									
D. nigra	0.38	10.8	26.2	42.3	20.8				
Subgenus *Engiscaptomyza* (Hawaiian species)									
D. crassifemur	0.29	10.6	21.2	38.6	29.6				
Subgenus *Scaptodrosophila*									
D. lebanonensis	0.45	5.6	31.8	41.3	21.4	33.7	28.0	18.0	20.4

A_4%, C_4%, T_4% and G_4% are the percentage of fourfold degenerate codons that end in A, C, T, or G, respectively; A_i%, C_i%, T_i% and G_i% are the percentage of A, C, T and G in introns, respectively.

[a] Only the sequences of the *Adh-2* gene have been determined.

19.4% C and 18.9% G. In synonymous sites, on the other hand, the A and C contents of highly biased genes were less than 10% and more than 50%, respectively. Similar tendencies were shown for all fourfold synonymous codon groups. Therefore, there seems to be strong negative selective constraints against A (and to a lesser extent T) and in favor of C at synonymous sites in highly biased genes. If tRNA abundance is a main cause of the selective constraints on synonymous sites, then perhaps there is a molecular reason for systematically excluding codons ending in U or A in preference for C, or perhaps the selective constraints at the translational level are unrelated to tRNA abundance. (We should also emphasize here that G contents do not show a correlation with codon-usage bias.)

It has been shown that substantial differences exist in level of expression and developmental pattern of *Adh* among Drosophila species (SULLIVAN, ATKINSON and STARMER 1990). In particular, *Adh* activity is quite different even between the closely related species *D. melanogaster* and *D. simulans*. However, Table 2 shows

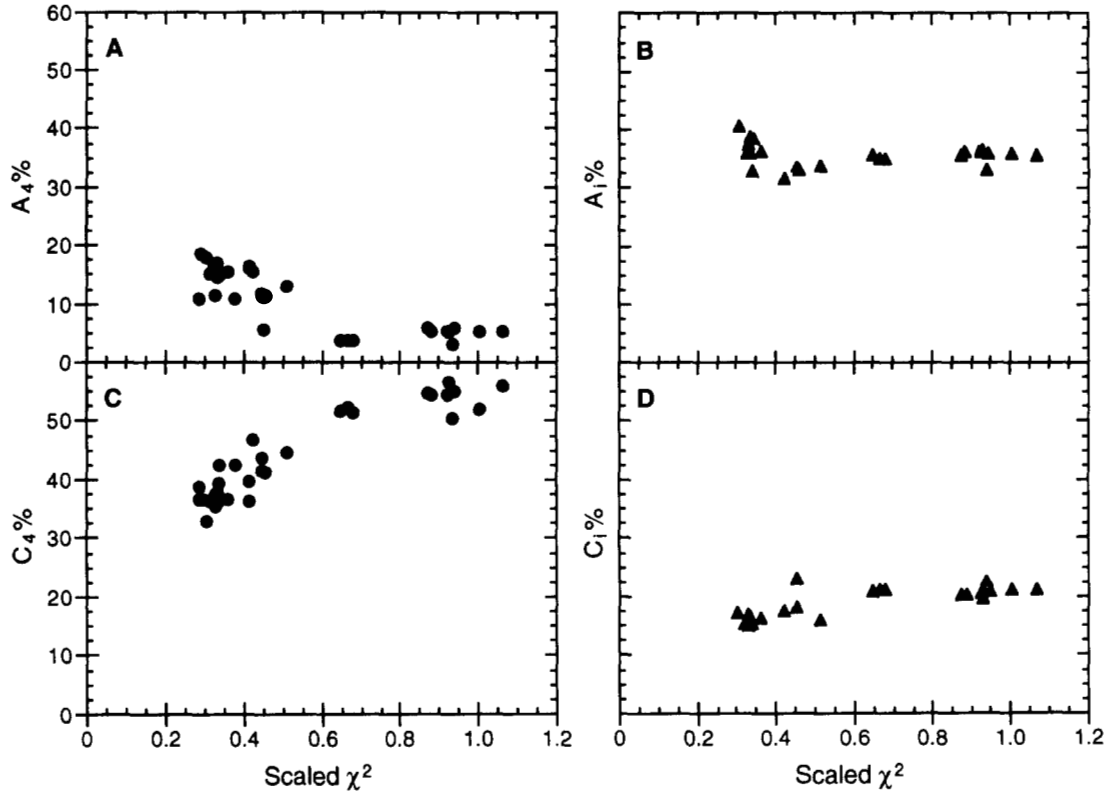

FIGURE 4.—Relation between scaled χ^2 and the identity of the third nucleotide in fourfold degenerate codons (A and C) and base composition of introns (B and D) for *Adh* genes of Drosophila. For species used, see MATERIALS AND METHODS and Table 2. A_4%, C_4%, A_i% and C_i% are defined as in Figure 1.

taxonomic changes in the scaled χ^2 and base composition at synonymous sites in the *Adh* genes among Drosophila lineages. Species within lineages are quite consistent, although there are some exceptions (*e.g.*, *D. ambigua* among the *obscura* group). The differences among lineages would appear to reflect differences in selective constraints on synonymous sites of the *Adh* genes during the evolution of the genus *Drosophila*.

Relative to the variation among lineages, we should also consider the effect of differences in generation time as mentioned earlier. For example, the *obscura* group has a generation time about twice as long as that of the *melanogaster* group, and the generation times of Hawaiian species are up to 10 times as long. However, because synonymous substitution rates correlate negatively with scaled χ^2 (SHARP and LI 1989; MORIYAMA and GOJOBORI 1992), and scaled χ^2 among Drosophila lineages correlates with base composition (Figure 4), the difference in generation time among Drosophila lineages does not seem to have a large effect on the variation in synonymous rate. In other words, synonymous sites of the Drosophila genome are not neutral enough to reflect an effect of generation time, if there is any.

Synonymous sites of Drosophila nuclear genes are evidently subject to some selective constraints. The selective constraints seem to occur post-transcriptionally, probably at the translational level, because the nucleotide sequences of introns greater than 500 bp do not seem to be subject to such selective constraints. It is possible that the degree of selective constraint for each gene simply reflects the level of expression of that particular gene. Some species of enterobacteria have a positive correlation between the level of gene expression and codon-usage bias (SHARP 1990, 1991). Some unicellular eukaryotes (yeasts and a slime mould) also seem to have the same feature (SHARP, TUOHY and MOSURSKI 1986; SHARP and WRIGHT 1988; SHARP and DEVINE 1989). If this is the case also in Drosophila, the synonymous sites of Drosophila seem to be highly sensitive to differences in level of expression among genes as well as sensitive to changes in the level of expression in divergent lineages. Because the synonymous substitution rate of a gene approximately doubles when the codon-usage bias decreases by half (SHARP and LI 1989; MORIYAMA and GOJOBORI 1992), the synonymous sites in the *Adh* gene of the *melanogaster* group may evolve at a rate less than half of that in the subgenus *Drosophila* as a whole (Table 2). For some other genes, however, the *melanogaster* group may have a higher rate than found in other species, because the level of gene expression should vary among genes and among species. Wide variation of synonymous substitution rates could affect the inference of phylogenetic trees and may serve as a caveat on methods that require homogeneity of rates across lineages.

We are grateful to JOHN CARULLI, DAN KRANE, STANLEY SAWYER and TAKASHI GOJOBORI (National Institute of Genetics at Mishima) for their valuable comments. We also thank JEFFREY POWELL who kindly provided the sequence data of *D. willistoni* before publication. This work was supported by a grant-in-aid from the Ministry of Education, Science and Culture, Japan (to E.N.M.) and by National Institutes of Health grant numbers GM40322 and GM33741 (to D.L.H.).

LITERATURE CITED

ANDERSON, C. L., E. A. CAREW and J. R. POWELL, 1993 Evolution of the *Adh* locus in the *Drosophila willistoni* group: the loss of an intron and shift in codon usage. Mol. Biol. Evol. (in press).

AOTA, S.-I., and T. IKEMURA, 1986 Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. **14:** 6345–6355.

ASHBURNER, M., and W. GELBART, 1991 *Drosophila* genetic maps. Drosophila Inform. Serv. **69**

BRITTEN, R. J., 1986 Rates of DNA sequence evolution differ between taxonomic groups. Science **231:** 1393–1398.

BULMER, M., 1987 A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol. Biol. Evol. **4:** 395–405.

BULMER, M., 1988 Codon usage and intragenic position. J. Theor. Biol. **133:** 67–71.

CARULLI, J. P., D. E. KRANE, D. L. HARTL and H. OCHMAN, 1993 Compositional heterogeneity and patterns of molecular evolution in Drosophila. Genetics (in press).

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

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

LAWRENCE, J. G., and D. L. HARTL, 1991 Unusual codon bias occurring within insertion sequences in *Escherichia coli.* Genetica **84:** 23–29.

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

LILJENSTRÖM, H., and G. VON HEIJNE, 1987 Translation rates modification by preferential codon usage: intragenic position effects. J. Theor. Biol. **124:** 43–55.

MARFANY, G., and R. GONZALEZ-DUARTE, 1991 The *Adh* genomic region of *Drosophila ambigua*: evolutionary trends in different species. J. Mol. Evol. **32:** 454–462.

MIYATA, T., T. YASUNAGA and T. NISHIDA, 1980 Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Natl. Acad. Sci. USA **77:** 7328–7332.

MORIYAMA, E. N., 1987 Higher rates of nucleotide substitution in Drosophila than in mammals. Jpn. J. Genet. **62:** 139–147.

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

SHARP, P. M., 1989 Evolution at "silent" sites in DNA, pp. 23–32 in *Evolution and Animal Breeding: Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson*, edited by W. G. HILL and T. F. C. MACKAY. C. A. B. International, Wallingford, U.K.

SHARP, P. M., 1990 Processes of genome evolution reflected by base frequency difference among *Serratia marcescens* genes. Mol. Microbiol. **4:** 119–122.

SHARP, P. M., 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33:** 23–33.

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:** 5029–5039.

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

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

SHARP, P. M., T. M. F. TUOHY and K. R. MOSURSKI, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5143.

SHARP, P. M., and F. WRIGHT, 1988 Analysis of yeast DNA sequence data: codon usage in the distantly related yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Yeast **4:** s515.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and W. WRIGHT, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SULLIVAN, D. T., P. W. ATKINSON and W. T. STARMER, 1990 Molecular evolution of the alcohol dehydrogenase genes in the genus Drosophila. Evol. Biol. **24:** 107–147.

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

Communicating editor: A. G. CLARK

APPENDIX

Nucleotide sequences of Drosophila and of *Adh* genes used in this study are given in Tables 3 and 4.

TABLE 3

Nucleotide sequences of Drosophila used in this study

Locus name	Accession no.	Cytological map position	Length of intron[a]	No. of codons	Scaled χ^2
D. melanogaster					
DROF16BPA	M76409	97A-B	1572	364	0.97
DROACT87EA	X12452	87E9-12	556	377	0.91
DROACT79B	M18829	79B		377	0.91
DROADHA	M14802	35B3	654	257	0.87
DROTUBA1	M14643	84B3-6		451	0.86
DRORPS17	M22142	67B1-5		132	0.80
DRODCO	X16969	30C1-6	906	354	0.73
DROHS83	X00065	63B-C	1140	375	0.71
DROPGD	M80598	2D6	1419	482	0.65
DROGPDHA	J04567	25F5	1196	351	0.60
DROMYLA	M11947	99E1-2	669	223	0.59
DROTUBA2	M14644	85E6-10		450	0.56
DROAFLL	M61127	72AB		181	0.56
DROMPP1, 2	M28418, M28999	44D-E		1287	0.53
DROAMA	M23561	84A		334	0.52
DROARR	M30177	36D1-2		365	0.52
DROEF1AF2	X06870	48D	1245	463	0.51
DROTUBA4	M14646	67C4-6		463	0.50
DROSODG	X13780	68A4-9	725	154	0.48
DROCHORS3	X05245	7F1-2		307	0.47
DROKNIRPS	X13331	77E1-2	733	430	0.45
DROLAMB2A	M58417	67C	2597	1640	0.38
DROUBX1, 2, 3	X05723, X05724, X05725	89E1-2		289	0.36
DRODMRP3	M62975	88A-B		260	0.36
DROGLASS	X15400	91A1-2	646	605	0.35
DROHBG	Y00274	85A3-B1		759	0.34
DROEIP28G	X04024	71C3-D2	996	256	0.34
DRODDC	X04661	37C1-2	1029	509	0.34
DROFASI	M32311	89E1-2	8747	653	0.34
DROL2AMD	X04695	37B13-C1		511	0.33
DROCCG	X04227	37C1-2		204	0.33
DRORPRIIA	M27431	10C2-D4	599	1897	0.33
DROH2AVDG	X15549	97C-D	545	141	0.32
DROHOXNK1	M27289	93E3-5		128	0.32
DROANTPG1, 2, 3, 4, 5	M14497, M14498, M14491, M14495, M14496	84A4-B2	2315	379	0.32
DROGLDGMC	M29298	84C8-D1	2983	613	0.32
DROTKABL1, 2, 3	M19690, M19691, M19692	73B1-2	2152	1521	0.31
DRODG2T1A3, 4, 5, 6	M27117, M27118, M27119, M27120	24A		1089	0.31
DROTRP	M34394	99C5-6		1276	0.30
DROARM	X54468	2B15		844	0.30
DROEGFRD	K03054	57F1	746	1284	0.29
DROSEV	J03158	10A1-2	5679	2555	0.29
DROBSG25D	X04896	25D3	1944	742	0.28
DRONOTCH1, 2, 3	M13689, K03507, K03508	3C7	8130	2704	0.26
DROXDH	Y00308	87D6-13	815	1336	0.26
DROPRD	M14548	33C1-2		614	0.26
DROGPAD	M31129	47A	1070	200	0.26
DRONANOS	M72421	91E-F	548	402	0.25
DROOSKAR	M65178	85B		607	0.25
DROTU36B	X15008	36B		415	0.24
DROBX200	X13168	89E1-2	1277	528	0.23
DROBCDG	X07870	84A1	1072	495	0.22
DROTOPII	X61209	37D2-6	933	1448	0.20
DROBOSS	X55887	96F8-11		896	0.18
DRODGQ	M58016	49B	515	354	0.18
DROPOLA	D90310			1479	0.17
DRONOONTA	M33496	14C1-2	1872	701	0.17
DROS2ZSTG	X56798		3949	1365	0.17
DROYELLOW	X04427	1B1	2719	542	0.16
DROOTUA	M30825	7F1	1118	812	0.16

TABLE 3

Continued

Locus name	Accession no.	Cytological map position	Length of intron[a]	No. of codons	Scaled χ^2
DROSUHW	Y00228	88B-C	552	945	0.15
DROLGL2	M17022	21A1-C2	1339	1162	0.12
DROSUSG	M57889	1B11-13	703	1335	0.11
DROPCXGEN	M74329	2E2-3	1744	2484	0.11
DROSCHSC1	X14396			281	0.08
DROKR	X03414	60F3		468	0.05
D. pseudoobscura					
DROUBXCA	X05179		1054	257	0.82
DROADHG	Y00602		794	255	0.68
DROHSP83	X03812		1063	375	0.63
DROGLDGMCA	M29299		1206	613	0.50
DROXDHA	M33977		1024	1343	0.49
DROPERAA	X13878		532	1242	0.34
DROGARTA	X06285		525	1365	0.32
D. virilis					
DROSEV1, 2, 3	M34543, M34544, M34545		7208	2595	0.65
DROHSP82N	X03811		955	374	0.50
DROHB	X15359		267	817	0.47
DROENGAA	X04727		2110	585	0.45
DROELAVG	M61748		1143	520	0.44
DROFMRFRN2	M32643		539	340	0.42
DROPERM	X13877		1421	1088	0.33
DROROUGHA	M35372		4513	340	0.24

[a] Total length for introns more than 500 bp (200 bp for *D. pseudoobscura* and *D. virilis*).

TABLE 4

Nucleotide sequences of *Adh* genes used in this study

Species	Locus name or reference	Accession no.	Length of intron
Subgenus *Sophophora*			
(*melanogaster* group)			
D. melanogaster	DROADHA	M14802	654
D. mauritiana	DROADHAAA	M19264	665
D. sechellia	DROADHG2	Z00045	653
D. simulans	DROADH01	X00607	656
D. erecta	DROADJ	X54116	614
D. orena	DROADHGY	M37837	619
D. yakuba	DROADK	X54120	583
D. teissieri	DROADL	X54118	633
(*obscura* group)			
D. pseudoobscura	DROADHG	Y00602	794
D. persimilis	DROPEADH17	M60997	769
D. miranda	DROMIRADH	M60998	766
D. subobscura	DROADHBB	M55545	764
D. ambigua	MARFANY and GONZALEZ-DUARTE (1991)		884
(*willistoni* group)			
D. willistoni	ANDERSON *et al.* (1993)		
Subgenus *Drosophila*			
(*repleta* group)			
D. mulleri	DROADHGZ	X03048	1993
D. arizonae[a]	DROADH2A	M62741	
D. mayaguana[a]	DROADH2B	M62742	
D. buzzatii[a]	DROADH2C	M62743	
D. wheeleri[a]	DROADH2D	M62851	
D. mojavensis	DROADHAAZ	M37276	3127
D. mettleri	DROADHM	M57300	1172
D. hydei	DROADH12P	X58694	1674
D. navojoa	DROADH1	X15585	1411
(Hawaiian picture-winged group)			
D. adiastola	DROADHZB	M60791	
D. picticornis	DROADHDPA	M63392	386
D. affinidisjuncta	DROADHAB	M37262	531
D. planitibia	DROADHDP	M63390	382
D. heteroneura	DROADHDH	M63287	381
D. silvestris	DROADHDS	M63291	382
D. differens	DROADHDD	M63303	383
(Hawaiian modified-mouthparts group)			
D. mimica	DROADHZC	M60792	
(Hawaiian fungus-feeders group)			
D. nigra	DROADHZD	M60793	
Subgenus *Engiscaptomyza* (Hawaiian species)			
D. crassifemur	DROADHZA	M60790	
Subgenus *Scaptodrosophila*			
D. lebanonensis	DROADHH	X54814	1281

[a] Only the sequences of the *Adh-2* gene have been determined.