

Natural Selection on Synonymous Sites Is Correlated With Gene Length and Recombination in *Drosophila*

Josep M. Comeron,* Martin Kreitman* and Montserrat Agudé†

*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and †Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain

Manuscript received July 6, 1998
Accepted for publication September 30, 1998

ABSTRACT

Evolutionary analysis of codon bias in *Drosophila* indicates that synonymous mutations are not neutral, but rather are subject to weak selection at the translation level. Here we show that the effectiveness of natural selection on synonymous sites is strongly correlated with the rate of recombination, in accord with the nearly neutral hypothesis. This correlation, however, is apparent only in genes encoding short proteins. Long coding regions have both a lower codon bias and higher synonymous substitution rates, suggesting that they are affected less efficiently by selection. Therefore, both the length of the coding region and the recombination rate modulate codon bias. In addition, the data indicate that selection coefficients for synonymous mutations must vary by a minimum of one or two orders of magnitude. Two hypotheses are proposed to explain the relationship among the coding region length, the codon bias, and the synonymous divergence and polymorphism levels across the range of recombination rates in *Drosophila*. The first hypothesis is that selection coefficients on synonymous mutations are inversely related to the total length of the coding region. The second hypothesis proposes that interference among synonymous mutations reduces the efficacy of selection on these mutations. We investigated this second hypothesis by carrying out forward simulations of weakly selected mutations in model populations. These simulations show that even with realistic recombination rates, this interference, which we call the "small-scale" Hill-Robertson effect, can have a moderately strong influence on codon bias.

GENES of flies, yeast, and bacteria show a great diversity of codon usage bias, and weak natural selection on synonymous mutations is generally believed to be the cause of this nonrandom usage (Grosjean and Fiers 1982; Sharp *et al.* 1986; Sharp and Li 1987; Shields *et al.* 1988). Also the rate of synonymous substitution (K_s) is inversely related to codon usage bias in both enterobacterial and *Drosophila* genes (Sharp and Li 1987, 1989), indicating a stronger selection as the cause of the more biased usage. In bacteria and yeast, the major (or preferred) codon always corresponds to the most abundant tRNA within each codon family (Bennetzen and Hall 1982; Ikemura and Ozeki 1983; Ikemura 1985). In these organisms, codon bias is also strongly correlated with the expression level of a gene (Grantham *et al.* 1981; Ikemura 1981; Bennetzen and Hall 1982; Gouy and Gautier 1982), indicating that selection is acting at the level of general translational efficiency (Sharp and Li 1987; Bulmer 1991). In *Drosophila*, preferred codons also appear to correspond to the more abundant tRNAs (White *et al.* 1973; Shields *et al.* 1988; Moriyama and Powell 1997), but there is specific evidence of selection for translational accuracy

rather than efficiency (Akashi 1994; Comeron and Kreitman 1998). In addition, although there is a qualitative positive relationship between expression levels and codon bias, there are also many counterexamples of highly expressed genes with little or no codon bias, suggesting that gene expression is not sufficient to explain levels of codon bias in *Drosophila* (Gonzalez *et al.* 1989; Fitch and Strausbaugh 1993). Therefore, while in unicellular organisms different codon bias among genes can be largely explained by different levels of expression, in *Drosophila* the variability of synonymous codon usage among genes remains unexplained.

Population genetics models of selection predict a positive relationship between the effectiveness of natural selection acting on a mutation and the recombination rate in the vicinity of the mutation. The Hill-Robertson effect (Hill and Robertson 1966; Felsenstein 1974) describes a general interaction between selection and recombination in finite populations. Under the Hill-Robertson effect the frequency and fixation probability of a given mutation are related not only to the selection coefficient of the specific mutation, but also to other mutations on the same genetic background, as well as to interference by linked mutations segregating in the population. Hitchhiking (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992) and background selection (Charlesworth *et al.* 1993; Charlesworth 1994) models, both of which have received con-

Corresponding author: Josep M. Comeron, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637. E-mail: jcomeron@midway.uchicago.edu

MATERIALS AND METHODS

siderable theoretical attention recently to account for nucleotide polymorphism levels in *Drosophila*, can be viewed as special cases of the more general Hill-Robertson effect. Under these models, a reduction in the effectiveness of selection in regions of low recombination, *i.e.*, tight linkage, can be interpreted as being equivalent to a reduction of the effective population size, N_e , for genes residing in these regions. A strong prediction, then, of these models is that neutral polymorphism levels, which are governed by the population size (Kimura 1983), will be positively correlated with recombination rates. This prediction has been confirmed in *Drosophila* by the otherwise inexplicable finding that polymorphism levels, but not evolutionary rates of sequence divergence, are correlated with recombination rates across the entire range of recombination rates (Begun and Aquadro 1992; Aguadé and Langley 1994; Aquadro *et al.* 1994).

The effectiveness of natural selection on slightly deleterious or advantageous mutations (Ohta and Kimura 1971) is also expected to be sensitive to N_e , and this implies that codon bias, if it is maintained by weak selection, will also be correlated with recombination rates. The correlation between codon bias and recombination rates, however, is detected only when genes in the regions of extremely restricted recombination are compared to genes in regions of high recombination rate (Kliman and Hey 1993). In these regions, nucleotide polymorphism levels, and hence N_e , differ by at least a factor of 10, suggesting that selection acting on synonymous mutations may be too strong to be nearly neutral. Contradicting this conclusion, however, is the more convincing observation that rates of synonymous evolution in the low codon bias genes residing in regions of extremely low recombination rate and in the high codon bias genes residing in regions of high recombination are not sufficiently different to be compatible with the strong selection hypothesis. Thus, the lack of a general correlation between recombination rates and codon bias levels in *Drosophila* implies that additional factors influencing codon bias must be operating in these species.

Recently, the length of coding regions in *Drosophila* genes has been shown to be positively correlated with the rate of synonymous substitution (Comeron and Aguadé 1996) and negatively correlated with the level of codon bias (Comeron 1997; Powell and Moriyama 1997). Here, we have analyzed the relationship among the length of the coding region, recombination rates, and codon bias in *Drosophila melanogaster*. We propose two models that account for the observed pattern of codon bias, divergence, and polymorphism, as the result of the interaction among the number of codons, translational efficiency, and recombination rates, in determining the effectiveness of selection on synonymous sites in *Drosophila*.

Sequences analyzed: Interactions among the length of the coding region, recombination rates, and synonymous codon usage bias were analyzed using 537 complete coding region sequences of *D. melanogaster* obtained from GenBank. Only those complete sequences located in specific polytene-chromosome map positions in FlyBase (1998; <http://flybase.bio.indiana.edu/>) were studied. Also, we discarded all multiple entries, multiple splicing products, and coding regions smaller than 100 codons to have accurate codon usage bias estimates. For families of alternatively spliced genes, we have included the longest one as a single representative; the number of such genes in the final set of 537 coding regions is 6. The complete list of the analyzed genes is available upon request. For the interspecific analysis, the number of synonymous (K_s) and nonsynonymous (K_a) substitutions per site was estimated for the 35 genes where homologous genes can be compared between *D. melanogaster* and at least one obscure group species (*D. subobscura* and *D. pseudoobscura*; see Comeron and Kreitman 1998, for details).

Codon usage analysis: Codon usage bias was estimated as the "effective number of codons" (ENC) (Wright 1990). This measure of the codon usage bias is independent of the number of codons under analysis and it exhibits a small index of dispersion (Comeron and Aguadé 1998). The two amino acids with only one codon (methionine and tryptophan) have not been taken into account to estimate the codon usage bias. Sixfold synonymous codons have been treated as two independent pools of two- and fourfold synonymous codons. Lower ENC values indicate a stronger codon usage bias, where the maximum codon usage bias produces an ENC estimate of 21, while the equifrequent usage of all synonymous codons gives an ENC value of 59.

Recombination rate: Recombination rates were estimated after obtaining the polynomial curves (Kliman and Hey 1993) as function of the quantity of DNA in each division along each chromosome (Sorsa 1988) *vs.* the change of the cytogenetic map position (FlyBase: Cytotable.txt). A single polynomial curve was obtained for the third chromosome ($R = 0.935$), while two independent curves were obtained for chromosomes *X* and *II* to better fit to the observed pattern of recombination rates ($R = 0.984$ and 0.918 for the *X* chromosome, and $R = 0.927$ and 0.760 for the second chromosome). Although recombination under natural conditions can be different than the standard map distances (see Charlesworth 1996), the possible error would be expected to obscure rather than enhance the observed patterns. A PC-Windows program to estimate the recombination rates in *D. melanogaster* based on the cytological map position using the described polynomial curves is available upon request to J.M.C.

Divergence estimates: The number of synonymous substitutions per site (K_s) was estimated as described in Comeron (1995), using the program K-estimator v4.0, available at <ftp.bio.indiana.edu/molbio/mswin/directory/or> from J.M.C.

Mutational matrix: A mutational matrix between nucleotides based on the observed polymorphisms at the noncoding region of the *Xdh* locus of *D. subobscura* was used (Comeron 1997). This mutational matrix predicts a G+C to A+T content at equilibrium of 40:60, and the expected random usage of synonymous codons based on this mutational pattern gives an ENC value of 58.0 and an average frequency of preferred codons of 0.352.

RESULTS

We analyzed the relationship between the length of the coding region and the rate of synonymous evolution

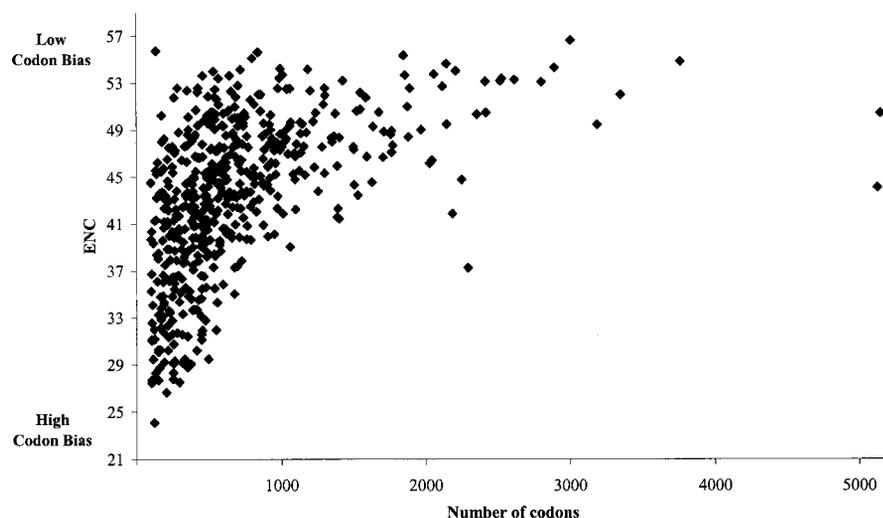


Figure 1.—Scatter plot of the codon bias (ENC) and the number of codons for 537 complete coding regions in *D. melanogaster*.

using 35 genes in which the comparison of the homologous sequences of *D. melanogaster* and at least one obscure species (*D. subobscura* and *D. pseudoobscura*) is possible (Cameron and Kreitman 1998). There is a positive relationship between the length of the complete coding region and K_s (nonparametric Kendall's test $\tau = 0.287$, $P = 0.017$; Sokal and Rohlf 1995). We then analyzed the relationship between the codon usage bias, measured as ENC, and the length of 537 complete coding regions in *D. melanogaster* (Figure 1). Codon bias is significantly reduced as a function of the length of the coding region ($\tau = 0.434$, $P < 0.0001$). Short genes exhibit a wide range of codon usage bias while long genes tend to be only slightly biased. Congruently, the length of the coding region is also correlated with base composition at the third position of fourfold degenerate codons, negatively with C_4 ($\tau = 0.265$, $P < 0.0001$) and positively with A_4 ($\tau = 0.220$, $P < 0.0001$) and T_4 ($\tau = 0.105$, $P < 0.001$); G_4 shows a small negative correlation ($\tau = 0.060$, $P = 0.039$). Equivalent results are obtained when the third position of all codons is used in the analyses (data not shown). The relationship between the length of the coding region and the synonymous base composition corroborates that there is no methodological bias associated with the estimation of ENC due to the number of codons under analysis.

We also analyzed the relationship between the estimated recombination rate and both the codon bias and the length of the coding region. A weak positive correlation between recombination rate and codon usage bias can be detected ($\tau = 0.057$, $P = 0.0465$) when all 537 genes are analyzed, while there is no relationship between length and recombination rates ($\tau = 0.01$, $P > 0.70$). Figure 2 shows the relationship among codon bias, recombination rate, and the length of the coding region for genes on the two large autosomes (chromosomes II and III); equivalent results are obtained for the X chromosome but for short genes (see discussion). The relationship between the recombination rate

and codon bias changes with the length of the coding region. For genes shorter than 750 bp there is a positive significant correlation between recombination rates and codon bias levels ($\tau = 0.168$, $P = 0.0163$), and this positive relationship is nearly linear across the entire range of recombination rates. The positive relationship between recombination rates and the codon bias levels becomes weaker for longer coding regions (for genes longer than 4500 bp, $\tau = 0.169$, $P = 0.0880$). For short genes, the correlation between recombination rates and codon bias levels remains significant ($\tau = 0.161$, $P = 0.0420$) even when genes located in regions of very low recombination rates—smaller than 10^{-10} /bp/generation (compared to an average recombination rate in *D. melanogaster* of 2×10^{-8} /bp/generation)—are excluded from the analysis. Also, a Kruskal-Wallis test (Sokal and Rohlf 1995) reveals higher average codon bias in regions with high recombination rates than in regions with low recombination rates ($H = 9.021$, $P = 0.011$, for the groups 1×10^{-10} – 1×10^{-9} , 1×10^{-9} – 1×10^{-8} , and 1 – 5×10^{-8} /bp/generation).

The length of a coding region is negatively correlated with codon bias across the entire range of recombination rates ($P < 0.0001$, for all cases). This negative correlation is also detected in those regions classified as non-recombining (recombination rates smaller than 10^{-10} /bp/generation) and exhibiting, on average, lower codon bias levels ($\tau = 0.532$, $P < 0.0001$). Even those genes located on the fourth chromosome (11 genes), a small autosomal chromosome that does not normally recombine (Hochman 1976), also show this significant correlation ($\tau = 0.600$, $P = 0.0102$). These correlations can be used to gauge the range of selection coefficients acting on synonymous mutations within a gene (see discussion).

A length-dependent selection coefficient (LdSC) model: Translational efficiency has been invoked as a general cause of codon bias in *Drosophila*. Taking the total time for a single ribosome to translate an mRNA

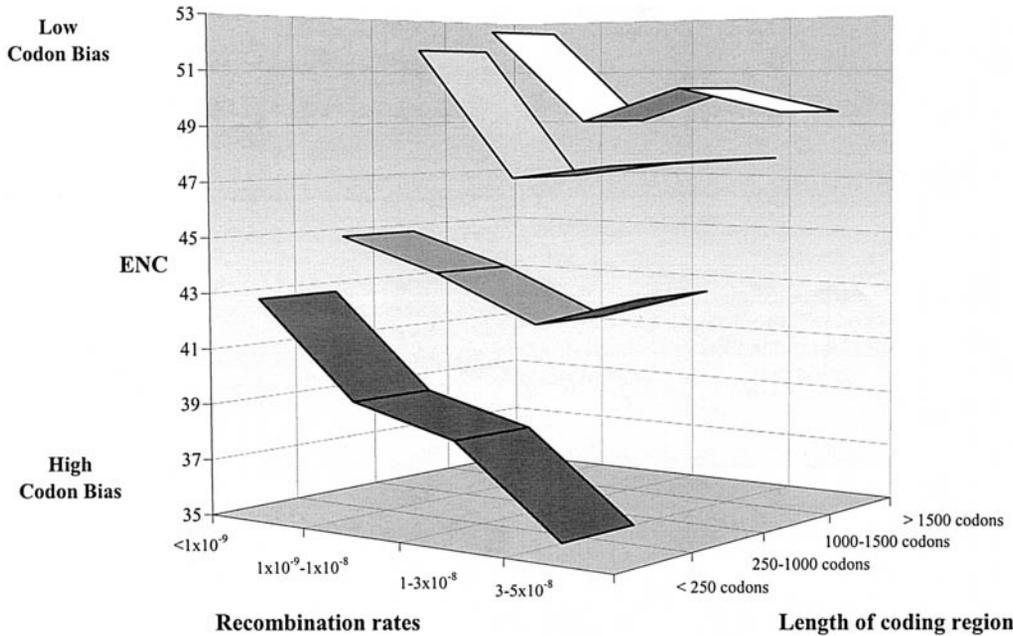


Figure 2.—Relationship among the length of coding region, the estimated recombination rate, and the codon bias (ENC) for the 426 complete genes of chromosomes II and III of *D. melanogaster*. The number of genes for the categories <250 codons, 250–1000 codons, 1000–1500 codons, and >1500 codons are 78, 278, 37, and 33, respectively.

as a proxy for translational efficiency (*i.e.*, fitness), consider the effect on translation time of a single synonymous mutation from an unpreferred to a preferred codon (preferred mutation). If the speed of translation of that amino acid increases by a fixed amount, then it is easy to show that its relative effect on the overall time of protein translation will be inversely related to the total number of codons of the coding region. The longer the coding region, the smaller the advantageous effect of a preferred mutation on the total translational time and, hence, on that protein's contribution to fitness.

We have investigated this simple model where the selection coefficient (s) on each synonymous mutation in a gene is inversely related to the number of codons (len) of a coding region. Let $s = s_{\max} len_{\min} / len$, where len_{\min} is the number of codons of the shortest coding region affected by different lengths (in our analyses $len_{\min} = 150$ codons; equivalent patterns are obtained using different len_{\min}), and s_{\max} the maximum value of s . The predicted codon bias (ENC) values can be obtained from the frequency of preferred codons (P ; Li 1987),

$$P = e^s V / (e^s V + U),$$

where for diploid organisms $S = 4N_e s$, $U = 4N_e u$, and $V = 4N_e v$, s is the selective advantage of the preferred codons, and u and v are the mutation rates from preferred to unpreferred and from unpreferred to preferred codons, respectively. It is assumed that $u + v \ll s$. All unpreferred codons are assumed to have the same fitness and their relative frequencies assumed to be those given by mutational equilibrium (see materials and methods). P_i and ENC_i were estimated separately for each i -fold ($i = 2, 3,$ and 4) synonymous group. Finally, $ENC = (12 ENC_2) + ENC_3 + (8 ENC_4)$.

Figure 3 shows the expected relationship between codon bias and coding region length predicted by the LdSC model. This model can explain much of the observed pattern of codon bias (see Figure 1) if one also assumes that genes have different maximum selection coefficients (s_{\max}), related to the specific contribution each protein makes to fitness.

Interference among synonymous mutations: Under the LdSC model, both the rate of synonymous divergence, K_s , and the level of synonymous polymorphism are negatively related to the level of codon bias. There is, in fact, a well-known negative correlation between K_s and codon bias among *Drosophila* species (Sharp and Li 1989). The level of synonymous polymorphism is, however, positively correlated with codon bias in *D. melanogaster* (Moriyama and Powell 1996), a correlation that is also significant when genes located in regions of no recombination are not used in the analysis ($\tau =$

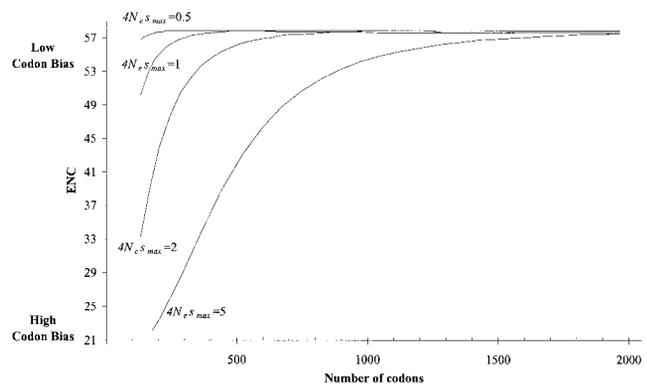


Figure 3.—Relationship between the number of codons and the predicted codon bias (ENC) based on the LdSC model (see text).

0.313, $P = 0.047$). Genes with low codon bias tend to have high rates of synonymous evolution but low levels of synonymous polymorphism. This disparity cannot be explained with a model assuming independent selection on synonymous mutations when two or more mutations are simultaneously segregating in a population, but it is directly predicted from a finding of Li (1987), who investigated the indirect effect (interference) among weakly selected mutations on codon bias in finite populations with no recombination.

To further assess the magnitude of the interference effect on codon bias, we investigated a Hill-Robertson-based model (see Introduction). The rationale was that genes with different lengths will have different numbers of synonymous mutations simultaneously segregating that can interfere with each other's fixation. We term this "finite population, mutation-selection, multisite, linkage model" as the small-scale Hill-Robertson (ssHR) model. If interference is stronger in longer genes than in shorter genes and of sufficient magnitude, then this model (Birky and Walsh 1988; Charlesworth *et al.* 1993; Charlesworth 1994) may be able to account for the empirical observations that longer genes have lower codon bias and higher rates of synonymous substitutions (K_s) in *Drosophila* and that genes with lower codon bias have lower levels of polymorphisms.

To study whether an ssHR model can shape the relationship between the synonymous codon usage and the coding region length, as observed in *Drosophila*, we carried out computer simulations to establish the codon bias at equilibrium for different coding lengths using plausible parameters of mutation, recombination, and selection on synonymous sites in *D. melanogaster*. Computer simulations conforming to the multisite model detailed in Li (1987) were performed with the following distinctions and features: (i) a diploid population of $2N$ nucleotide sequences each with L sites; (ii) each sequence represents a coding sequence of synonymous codon positions, basically equivalent to the third codon positions, assuming an equivalent usage of amino acids; (iii) each pool of synonymous codons has one preferred codon, and all unpreferred codons have a deleterious selection coefficient of s ; (iv) mutations are semidominant and the fitness is multiplicative over sites; (v) both the number of mutations and recombination events are Poisson distributed with a mean of $2NL\mu$ and $NL3c$, respectively, with μ the mutation rate per base pair per generation and c the recombination rate per base pair per generation; (vi) each generation is randomly obtained from the previous generation on the basis of the relative fitness of the diploid individuals; and (vii) substitution among nucleotides follows a previously estimated mutation matrix (see materials and methods). Computer simulations were conducted using the following parameters: $N = 1000$; $4N_s = 1.0, 2.0, 4.0, \text{ and } 10.0$; $2N\mu = 10^{-2}$; and $2Nc = 0, 0.04, \text{ and } 0.4$. The average codon bias (ENC) was obtained by sampling a minimum

of 10,000 generations after the population shows no directional change in either ENC or heterozygosity levels. Figure 4 shows that the effect of the intragenic interference among synonymous mutations is observed even with the maximum recombination rate in *D. melanogaster*. Also, for a given selection coefficient and recombination rate, long genes tend to show lower codon bias, except for $4N_s = 1.0$ where this pattern is detected only for sequences longer than 500 codons. Synonymous polymorphism levels also decrease conspicuously with gene length for $4N_s$ of 4, 2, and 1, and high recombination, where nucleotide diversity (π) for coding regions of 1500 codons shows a reduction of 25, 20, and 26%, respectively, compared to genes of 100 codons. For plausible recombination rates, this interference among sites will be restricted within a given gene. Nevertheless, in the extreme case of regions with very low recombination rates, or genes very close to each other, synonymous sites among linked genes will begin to interfere with each other; in those regions, the length and number of genes in the region as well as the magnitude of linkage will contribute to the bias in each gene.

Comparison among low-recombining regions of *D. melanogaster*: Genes located on the fourth chromosome in *D. melanogaster*, as in any other nonrecombining region, are expected to exhibit a smaller effect of natural selection on synonymous sites (Berry *et al.* 1991; Kreitman and Antezana 1998). Indeed, the fourth chromosome exhibits on average a significantly lower codon bias when compared to both the other autosomes ($P < 0.0001$, applying the nonparametric Mann-Whitney test) and the *X* chromosome ($P < 0.0001$). More interesting, this difference is also significant ($P = 0.0084$) when the fourth chromosome is compared to those genes located in regions of very low recombination rates (smaller than 1×10^{-10} /bp/generation) in the other autosomes (ENC of 50.91 and 42.60, for the fourth and the other autosomes, respectively). An independent observation indicates that the fourth chromosome has a number of transposable elements (TEs) equivalent to those regions with low recombination in other autosomes [$\chi^2 = 0.22, P = 0.64$, using data from Charlesworth *et al.* (1992)]. The abundance of TEs is thought to be basically regulated by the deleterious consequences on fitness of recombination between non-homologous elements (Montgomery *et al.* 1987; Langley *et al.* 1988; Charlesworth *et al.* 1992) and then negatively correlated with the recombination rate. Conversely, abundance of TEs is expected to be independent of gene conversion rates. Our observations, then, can hardly be explained by different, although residual, recombination rates between these regions of the chromosomes *II* and *III* and the fourth chromosome. A stronger effect of background selection or hitchhiking in the fourth chromosome based on a higher density of genes seems also unlikely as the average density is actually smaller in the fourth chromosome than in these

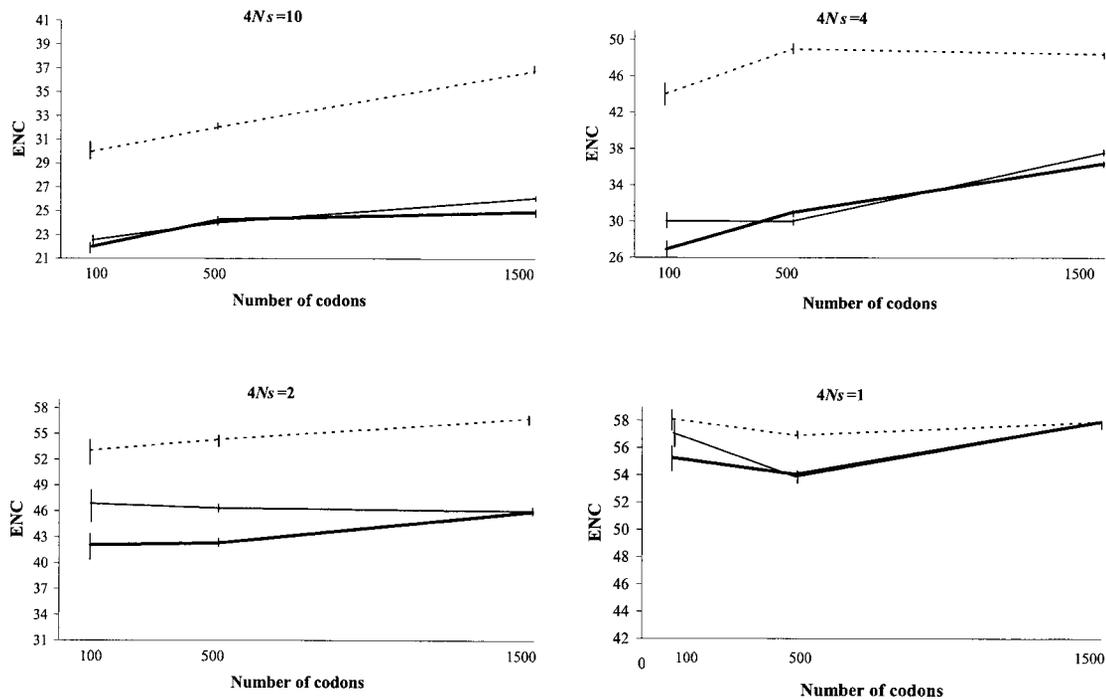


Figure 4.—Relationship between the length of the coding region and the codon bias obtained by computer simulation based on the ssHR model (see text) for coding regions of 300, 1500, and 4500 bp. Thick lines, thin lines, and dashed lines indicate $2Nc = 0.4$, $2Nc = 0.04$, and $2Nc = 0$, respectively. The expected ENC values with free recombination (Li 1987) are 21.00, 22.73, 34.14, and 50.89 for $4N_s$ of 10, 4, 2, and 1, respectively.

regions of the other autosomes (1.04 and 1.43 genes/band, respectively). Two causes, in principle, can be invoked to explain the data. First, some genes located in regions of very low recombination of the chromosomes *II* and *III* in *D. melanogaster* have not been placed in this recombinational environment for a long enough time to have reached the new equilibrium of synonymous base composition, a situation that is unlikely for the genes on the fourth chromosome. Second, rates of gene conversion higher in the regions of very low recombination in the chromosomes *II* and *III* than in chromosome *IV* could also explain the observations. In fact, differential exchange of mutations among chromosomes has already been detected in different regions with very low recombination of the *X* chromosome of *D. melanogaster* (Begun and Aquadro 1995). Large-scale polymorphism studies in these regions will be needed to test this last prediction.

DISCUSSION

Effect of the length of the coding region and recombination rates on codon bias and synonymous substitutions rates: We have confirmed the positive relationship between the length of the coding region and the rate of synonymous substitutions per site (K_s) in *Drosophila*. We have also shown that there is a negative correlation between the length of the coding region and the codon usage bias in *D. melanogaster*. Together, both results indi-

cate that the strength of selection on synonymous mutations, or the effectiveness of selection, is on average smaller in long genes than in short ones. The length of the coding region is also correlated with the base composition at synonymous sites in the sense that very long genes show a base composition closer to that expected under the mutational equilibrium (see Table 1), which can be inferred from the nucleotide composition of long introns under the assumption of neutrality (Moriyama and Hartl 1993). Thus, higher selection pressures to achieve and maintain a base composition markedly different from that expected under mutational equilibrium, and hence a high codon bias, explain the lower estimates of K_s for shorter coding regions.

The observed nonrandom usage of the different synonymous codons in most organisms indicates that synonymous mutations are not strictly neutral, but rather subject to weak selection. The nearly neutral theory of molecular evolution proposes that the magnitude of most selective coefficients on synonymous mutations should be of the order of the reciprocal of the effective population size (N_e). In *Drosophila* synonymous mutations are often assumed to behave nearly neutrally even when the data do not fully support that argument. Recombination rates positively correlate with the level of polymorphism in *D. melanogaster* (Begun and Aquadro 1992), indicating differences in the regional N_e . Nevertheless, the predicted effect of recombination on codon bias is observed only in regions with very low recombina-

TABLE 1
Average synonymous base composition in fourfold degenerate codons and codon bias (ENC)
compared to base composition in noncoding sequences

	<i>n</i>	A	G	C	T	ENC
Coding region						
Short genes (<750 bp)	98	0.170	0.231	0.423	0.176	37.82
Long genes (>4500 bp)	48	0.226	0.242	0.333	0.199	49.75
High recombination (HR)	208	0.180	0.261	0.402	0.157	43.27
Low recombination (LR)	42	0.237	0.218	0.330	0.215	46.41
HR and short genes	33	0.152	0.253	0.446	0.149	35.98
LR and long genes	4	0.315	0.182	0.223	0.280	53.36
Noncoding region ^a		0.301	0.189	0.194	0.314	55.99 ^b

High and low recombination regions are defined as $>3 \times 10^{-8}$ /bp/generation and $<1 \times 10^{-9}$ /bp/generation, respectively.

^a Base composition of the nontranscribed strand of long introns (from Moriyama and Hartl 1993).

^b Expected ENC measure assuming nucleotide frequencies at the third position of codons equivalent to those of the noncoding region and random amino acid usage.

tion (Kl iman and Hey 1993). Thus, the lack of a general correlation between the codon bias and recombination suggests that selection on most synonymous mutations might be too large, challenging the weak selection hypothesis.

Here we have shown that recombination rates correlate with the effectiveness of selection on synonymous mutations, and hence also with codon bias, even when genes located in regions with very low recombination are excluded from the analyses. This correlation, however, is also dependent on the length of the coding region, as indicated by the systematic shift in codon bias levels with recombination rates in short genes but not, or to a lesser degree, in long genes (Figure 2). The recombination environment of a gene does, therefore, affect the evolutionary behavior of synonymous mutations, as predicted by the weak selection theory.

Shuffling of gene order within chromosome arms, a common feature of *Drosophila* evolution (Pinsker and Sperlich 1984; Whiting *et al.* 1989; Segarra and Aguadé 1992; Kress 1993; Segarra *et al.* 1996), is expected to mediate the effectiveness of selection on synonymous mutations through changes in the recombinational environment of a gene (Charlesworth 1994). A change of codon bias is then predicted in those genes in regions that have changed their recombination rates, as observed for the *yellow* and *scute* genes between *D. melanogaster* and *D. subobscura* (Munté *et al.* 1997). This may also explain why the variance in the rate of synonymous evolution for *Drosophila* genes is significantly greater than expected under the neutral model of evolution (Kimura 1983; Gillespie 1989). Zeng *et al.* (1998) have shown that 13 out of the 24 analyzed genes show a significantly high variance of synonymous substitutions rates among the *D. melanogaster*, *D. subobscura*, and *D. pseudoobscura* lineages. The extensive gene reorder within chromosomal arms observed between *D. melanogaster* and the *obscura* species (Pinsker and Sperlich

1984; Segarra and Aguadé 1992; Segarra *et al.* 1996) could account for the change of substitution rates mainly due to the *D. melanogaster* lineage.

Why is codon bias systematically lower in larger genes? Selection on synonymous sites has been proposed to act, at least, at three different levels: the translational accuracy and translational efficiency levels and the mRNA level. Selection acting at the level of translational accuracy has been detected in *Drosophila* with two independent approaches: (1) regions encoding important protein motifs present a significantly higher codon bias (Akashi 1994) and (2) there is an excess of codons with nonindependent synonymous and nonsynonymous substitutions, not explainable by adjacent mutations, as shown by a phylogenetic analysis of *D. subobscura*, *D. pseudoobscura*, and *D. melanogaster* sequences (Comeron and Kreitman 1998). In these studies, however, selection acting at this level has been restricted to particular regions or codon positions, respectively, not associated with gene length. Also, a negative relationship between the length of the coding region and the codon bias due to accuracy of translation supports the idea that selection acts less efficiently in long genes than in short ones. If correct, this leads to the prediction that the rate of nonsynonymous substitutions (K_a) will be smaller in short genes than in long genes. There is, however, no correlation in *Drosophila* between the length of the coding region and the K_a estimates ($\tau = 0.048$, $P > 0.65$) for the 35 analyzed genes (Comeron and Kreitman 1998).

Elongation rates affect the overall speed of translation because the most abundant tRNA translates its corresponding codon faster (Andersson and Kurland 1990; Bulmer 1991). We have assumed selection acting to maintain translational efficiency linked to the speed of translation; the relative effect of a change between preferred and unpreferred codons on the total time of translation is smaller for longer genes. We have ex-

plored a simple LdSC model, where selection coefficients on individual synonymous mutations are different for different genes; the longer the gene, the smaller the coefficient. As a direct consequence, the length of the coding region will be positively correlated with the rate of synonymous substitutions per site and negatively correlated with the codon usage bias (Figure 3). The complete scenario would also incorporate differential contributions to fitness from different coding regions, represented in our model by s_{\max} (Figure 3). Assuming a minimum value of ENC (maximum codon bias) of 27–28 in *D. melanogaster* (see Figure 1) the highest estimate of $4N_e s_{\max}$ would be 2.7. This estimate however would be an underestimation if interference among synonymous mutations is actually playing a detectable role.

An alternative scenario based on translational efficiency also predicts that the level of gene expression will be inversely related to the length of the coding region. This is also an LdSC model because the selection coefficients for individual mutations will be negatively correlated with gene length. This possibility, however, seems unlikely in *Drosophila* because it would require the assumption that highly expressed genes have eliminated nonessential amino acids by shortening their coding length in order to improve translational efficiency. In that case, short genes should have lower rates of nonsynonymous substitution (K_a), which is not the case (see above).

Selection pressure at the mRNA level to maintain the stability of mRNA secondary structure has been proposed to have secondary or modulating effects on synonymous base composition (Bulmer 1991). In the regions, or sites, of a coding region affected by selection at the mRNA level, conflicting selection pressures are expected between optimizing base composition to stabilize mRNA molecules and biasing codon usage to improve the efficiency of translation. The smaller selection coefficients at the level of translation in long genes should allow a better optimization of mRNA structure, and its detection, in these genes in *Drosophila* (Comeron and Aguadé 1996).

The observed relationship between the length of the coding region and the codon bias in nonrecombining regions provides support for an LdSC model. In these regions, the area affected by selective sweeps (hitchhiking) or by the constant removal of a fraction of chromosomes with deleterious mutations (background selection) will be expected to contain genes of different lengths. If this is the case, codon bias levels will vary with gene length *only* if in these genes the selection coefficients on individual mutation also vary with length.

Magnitude of selection coefficients on synonymous sites in *Drosophila*: The effect of selection is detectable only when the product of N_e and the selection coefficient (s) is equivalent or larger than unity ($|N_e s| \geq 1$) (Kimura 1983; Li 1978). Nucleotide polymorphism lev-

els, and hence N_e , in regions with no detectable recombination in *D. melanogaster* have been estimated to be one (Begun and Aquadro 1992; Aguadé and Langley 1994; Aquadro *et al.* 1994) or two [in the extreme case of the fourth chromosome (Berry *et al.* 1991)] orders of magnitude smaller than the average in *D. melanogaster*. This being the case, the existence of a significant relationship between codon bias and the length of the coding region for genes located in regions of high recombination rates and also for those genes located in regions with no detectable recombination, such as the fourth chromosome, is clear evidence that a fraction of synonymous mutations must have selection coefficients large enough to satisfy the condition $|N_e s| \geq 1$. Also, neither the very long coding sequences nor the genes located in regions with very low recombination actually attain the same base composition as introns or a random usage of synonymous codons (Table 1). Only the combination of both causes of reduction of $N_e s$, smaller selection coefficients per mutation (s) in long genes and smaller $N_e s$ in regions of low recombination, gives rise to a synonymous base composition close to that expected under mutational equilibrium. The distribution of the selection coefficients on synonymous mutations appears to be, therefore, extremely wide in *Drosophila*, encompassing one or two orders of magnitude, since their effect is detected across the entire range of recombination rates. Models of weak selection pertaining to the evolution and maintenance of codon bias need to be revised to reflect variable selection coefficients.

There is some controversy about whether codon bias selection acts more strongly through its action on advantageous (preferred) or deleterious (unpreferred) mutations. A difference in the efficacy of selection on the two types of mutations can occur if they are not semi-dominant. Comparison of *X*-linked and autosomal genes can shed some light on this discussion. In *Drosophila*, genes on the *X* chromosome are hemizygous in males; selection is expected to eliminate deleterious mutations on this chromosome more efficiently than on autosomes if mutations are partially or fully recessive. Then, for equivalent recombination rates, the proportion of alleles free of deleterious mutations (f_0) is expected to be larger for the *X* chromosome than for the autosomes (Charlesworth *et al.* 1993; Aquadro *et al.* 1994; Charlesworth 1994). In this case N_e will be approximately equal to f_0 , and so the effectiveness of selection on slightly neutral mutations should increase with f_0 . Because N_e for the *X* chromosome is 3/4 that expected for autosomes, this prediction is conservative. For our purposes, *X*-linked genes will be expected to have higher codon bias, on average, than autosomal genes. The opposite behavior is expected for advantageous mutations that are partially or fully recessive: codon bias will be lower, on average, for *X*-linked genes than for autosomal genes because selective sweeps would reduce N_e and the effectiveness of selection. We

TABLE 2
Observed relationships in *D. melanogaster* and predictions of the models

Factor	Observed relationship in <i>D. melanogaster</i>	Mechanism governing codon bias: predictions		
		Translational accuracy ^a	Translational time (LdSC) ^b	Local interference (ssHR) ^b
Codon bias	Decreasing with length ^c	Increasing	Decreasing	Decreasing
Codon bias	Decreasing with length in regions of low recombination	Increasing	Decreasing	Independent
Synonymous substitutions (K_s)	Increasing with length	Decreasing	Increasing	Increasing
Nonsynonymous substitutions (K_a)	No relationship with length	Decreasing	No relationship	No relationship–Increasing ^f
Efficacy of selection on synonymous mutations $ N_e s $	Decreasing with length ^d	Increasing	Decreasing	Decreasing
Synonymous polymorphism	Increasing with codon bias ^e	Decreasing	Decreasing	Increasing ^g

^a Eyre-Walker (1996a).

^b See text for details.

^c Total length of the coding region.

^d Inferred from patterns of codon bias and K_a .

^e Moriyama and Powell (1996).

^f Conditioned by the divergence time and selection coefficients on nonsynonymous mutations.

^g For a given selection coefficient on synonymous mutations (s).

focused our comparison on short genes, because codon bias in these genes is more clearly influenced by selection. For genes smaller than 750 bp, codon bias is higher for X-linked genes than for the second and third chromosome genes (ENC of 35.58 and 38.14, respectively), suggesting partial recessivity of unpreferred mutations. Congruent with this interpretation, the statistically significant positive relationship between the recombination rate and codon bias detected for autosomal genes smaller than 750 bp ($\tau = 0.185$, $P = 0.0163$) is not detected for X-linked genes ($\tau = 0.198$, $P = 0.25$, and in the opposite direction). Very long genes (longer than 4.5 kb), however, show equivalent codon bias in X-linked and autosomal genes. Both observations are consistent with the fact that short genes, which tend to have a higher codon bias, and therefore a higher frequency of preferred codons, will have a high proportion of unpreferred mutations. If these unpreferred mutations are partially recessive, they will be more effectively selected against if they are in X-linked genes, and we conjecture that this can lead to a greater independence of selection in relation to the X chromosome recombinational environment. This effect will be less pronounced in long genes, which will have roughly equivalent frequencies of deleterious and advantageous mutations.

In *Escherichia coli*, codon bias is positively correlated with gene length for equivalently expressed genes (ribosomal proteins); longer genes tend to have higher codon bias than the shorter ones (Eyre-Walker 1996a).

This relationship is the opposite of the one observed in *Drosophila* (Figure 1). Eyre-Walker and Bulmer (1993) also showed in *E. coli*, however, that genes with high codon bias exhibited significantly lower codon bias in the 5' end of the gene (up to 100 codons) than in the remainder of the gene. They attributed this reduced codon bias in the 5' region of genes to selective constraints on third positions of these codons imposed by ribosome assembly and translation initiation requirements. A similar trend has also been reported in amino-terminal codons as a result of the overlap of coding regions of many *E. coli* genes (Eyre-Walker 1996b). To take this effect into account, we have analyzed 1218 *E. coli* genes >250 codons in length (ECD Release 28; Wahl *et al.* 1994). When the first 100 and last 50 codons are removed, we find no significant relationship between codon bias and gene length (data available upon request). The lack of any relationship between gene length and codon bias (taking initiation constraints into account) in *E. coli* is not an unexpected result. Bulmer (1991), for example, argued that translation initiation (and not elongation rate) is rate limiting in *E. coli* protein synthesis. With this as an assumption, his analytical model of codon bias selection clearly shows that there is no relationship predicted between codon bias and gene length. The fact that such a relationship exists in *Drosophila* strongly suggests that translation initiation is not rate limiting in this species. Thus, although certain other features of codon bias are similar in both *E. coli*

and *Drosophila*, such as positive relationships between codon bias and gene expression levels and between tRNA abundance and codon preference, the actual mechanism governing codon selection may be different in the two species. In *E. coli*, the selective advantage of an increase in translation speed resulting from a mutation to a preferred codon accrues indirectly through its contributions to free ribosome pools. In our model the selective advantage of a preferred mutation accrues directly through increased production of that gene's product.

In conclusion, the evidences presented here confirm the importance of natural selection in shaping patterns of codon bias in *Drosophila*. The action of natural selection can be understood only as a complex interaction of several factors, including the length of the coding region, recombination rates, and variable selection strengths. While an LdSC model is sufficient to explain features of the relationship between codon bias and coding region length, even in regions with no detectable recombination, only an ssHR model can explain why genes with low codon bias show low levels of synonymous polymorphism (see Table 2). Both models predict that genes with low codon bias will be longer on average and will have higher levels of synonymous divergence than genes with high codon bias. The positive relationship between polymorphism and codon bias, however, is tentative and must be confirmed with a larger number of randomly chosen genes, taking into account both the recombination rate and the length of the coding regions. Testing this specific prediction of the ssHR model is important because standard models of molecular variation and evolution, and the statistical tests that derive from these models, assume that selection acts independently across mutations. Our analysis suggests that this may not be the case for synonymous mutations.

We thank P. Andolfatto, M. Antezana, C. Bergman, J. Braverman, B. Charlesworth, R. R. Hudson, A. Llopart, E. Stahl, W. Stephan, C. Toomajian, and M. Przeworski for helpful comments on the manuscript and A. Llopart for many useful discussions. We also thank B. Chen for his help compiling the data set of complete genes in *D. melanogaster*. J.M.C. is supported by a Postdoctoral Fellowship from Ministerio de Educación y Ciencia, Spain. This work was supported by a National Institutes of Health grant GM-39355 to M.K.

LITERATURE CITED

- Aguadé, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination in *Drosophila*, pp. 67–76 in *Non-neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.
- Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**: 927–935.
- Andersson, S. G., and C. G. Kurland, 1990 Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54**: 198–210.
- Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and the levels of DNA polymorphism in *Drosophila*, pp. 45–56 in *Non-neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.
- Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- Begun, D. J., and C. F. Aquadro, 1995 Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. *Mol. Biol. Evol.* **12**: 382–390.
- Bennetzen, J. L., and B. D. Hall, 1982 Codon selection in yeast. *J. Biol. Chem.* **257**: 3026–3031.
- Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129**: 1111–1117.
- Birky, C. W., Jr., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**: 6414–6418.
- Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–907.
- Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**: 213–227.
- Charlesworth, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* **68**: 131–149.
- Charlesworth, B., A. Lapid and D. Canada, 1992 The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II. Inferences on the nature of selection against elements. *Genet. Res.* **60**: 115–130.
- Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- Comeron, J. M., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* **41**: 1152–1159.
- Comeron, J. M., 1997 Estudi de la variabilitat nucleotídica a *Drosophila*: regio *Xdh* a *D. subobscura*. Ph.D. Thesis, Universitat de Barcelona, Barcelona, Spain.
- Comeron, J. M., and M. Aguadé, 1996 Synonymous substitutions in the *Xdh* gene of *Drosophila*: heterogeneous distribution along the coding region. *Genetics* **144**: 1053–1062.
- Comeron, J. M., and M. Aguadé, 1998 An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* **47**: 268–274.
- Comeron, J. M., and M. Kreitman, 1998 The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection, or relaxed constraints? *Genetics* **150**: 767–775.
- Eyre-Walker, A., 1996a Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol. Biol. Evol.* **13**: 864–872.
- Eyre-Walker, A., 1996b The close proximity of *Escherichia coli* genes: consequences for Stop codon and synonymous codon use. *J. Mol. Evol.* **42**: 73–78.
- Eyre-Walker, A., and M. Bulmer, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* **21**: 4599–4603.
- Felsenstein, J., 1974 The evolutionary advantage of recombination. *Genetics* **78**: 737–756.
- Fitch, D. H., and L. D. Strausbaugh, 1993 Low codon bias and high rates of synonymous substitution in *Drosophila hydei* and *D. melanogaster* histone genes. *Mol. Biol. Evol.* **10**: 397–413.
- FlyBase, 1998 FlyBase—a *Drosophila* database. *Nucleic Acids Res.* **26**: 85–88.
- Gillespie, J. H., 1989 Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* **6**: 636–647.
- Gonzalez, F., S. Romani, P. Cubas, J. Modolell and S. Campuzano, 1989 Molecular analysis of the *asense* gene, a member of the *achaete-scute* complex of *Drosophila melanogaster*, and its novel role in optic lobe development. *EMBO J.* **8**: 3553–3562.
- Gouy, M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**: 7055–7074.
- Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**: R43–74.
- Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**: 199–209.
- Hill, W. G., and A. Robertson, 1966 The effect of linkage on the limits to artificial selection. *Genet. Res.* **8**: 269–294.
- Hochman, B., 1976 The fourth chromosome of *Drosophila melanogaster*.

- ter, pp. 903–928 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner and E. Novitski. Academic Press, London.
- 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* translational system. *J. Mol. Biol.* **151**: 389–409.
- Ikemura, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13–34.
- Ikemura, T., and H. Ozeki, 1983 Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harbor Symp. Quant. Biol.* **47**: 1087–1097.
- Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The “hitchhiking effect” revisited. *Genetics* **123**: 887–899.
- Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**: 1239–1258.
- Kreitman, M., and M. Antezana, 1998 The population and evolutionary genetics of codon bias, in *Evolutionary Genetics from Molecules to Morphology*, edited by R. Sing and C. Krimbas. Columbia University Press, New York (in press).
- Kress, H., 1993 The salivary gland chromosomes of *Drosophila virilis*: a cytological map, pattern of transcription and aspects of chromosome evolution. *Chromosoma* **102**: 734–742.
- Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan and B. Charlesworth, 1988 On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52**: 223–235.
- Li, W.-H., 1978 Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics* **90**: 349–382.
- Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**: 337–345.
- Maynard Smith, J., and J. Haigh, 1974 The hitchhiking effect of a favorable gene. *Genet. Res.* **23**: 23–35.
- Montgomery, E., B. Charlesworth and C. H. Langley, 1987 A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet. Res.* **49**: 31–41.
- Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* **134**: 847–858.
- Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**: 261–277.
- Moriyama, E. N., and J. R. Powell, 1997 Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* **45**: 514–523.
- Munté, A., M. Aguadé and C. Segarra, 1997 Divergence of the yellow gene between *Drosophila melanogaster* and *D. subobscura*: recombination rate, codon bias and synonymous substitutions. *Genetics* **147**: 165–175.
- Ohta, T., and M. Kimura, 1971 On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1**: 18–25.
- Pinsker, W., and D. Sperlich, 1984 Cytogenetic mapping of enzyme loci on chromosomes *J* and *U* of *Drosophila subobscura*. *Genetics* **108**: 913–926.
- Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **94**: 7784–7790.
- Segarra, C., and M. Aguadé, 1992 Molecular organization of the X chromosome in different species of the obscura group of *Drosophila*. *Genetics* **130**: 513–521.
- Segarra, C., G. Ribó and M. Aguadé, 1996 Differentiation of Muller’s chromosomal elements D and E in the obscura group of *Drosophila*. *Genetics* **144**: 139–146.
- Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**: 222–230.
- Sharp, P. M., and W.-H. Li, 1989 On the rate of DNA sequence evolution in *Drosophila*. *J. Mol. Evol.* **28**: 398–402.
- Sharp, P. M., T. M. F. Tuohy and K. R. Mosurski, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**: 5125–5139.
- Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**: 704–716.
- Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*, Ed. 3. W. H. Freeman, New York.
- Sorsa, V., 1988 *Chromosome Maps of Drosophila*. CRC Press, Boca Raton, FL.
- Stephan, W., T. H. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* **41**: 237–254.
- Wahl, R., P. Rice, C. M. Rice and M. Kroger, 1994 ECD—a totally integrated database of *Escherichia coli* K12. *Nucleic Acids Res.* **22**: 3450–3455.
- White, B. N., G. M. Tener, J. Holden and D. T. Suzuki, 1973 Analysis of tRNAs during the development of *Drosophila*. *Dev. Biol.* **33**: 185–195.
- Whiting, J. H. Jr., M. D. Philey, J. L. Farmer and D. E. Jeffery, 1989 *In situ* hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis*. *Genetics* **122**: 99–109.
- Wright, F., 1990 The “effective number of codons” used in a gene. *Gene* **87**: 23–29.
- Zeng, L.-W., J. M. Comeron, B. Chen and M. Kreitman, 1998 The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila*. *Genetica* **102/103**: 369–382.

Communicating editor: A. G. Clark