Interactions Between Natural Selection, Recombination and Gene Density in the Genes of Drosophila

Jody Hey*,[1] and Richard M. Kliman†

Department of Genetics, Rutgers University, Piscataway, New Jersey 08854-8082 and †Department of Biological Sciences, Kean University, Union, New Jersey 07083

Manuscript received August 14, 2001
Accepted for publication November 19, 2001

ABSTRACT

In Drosophila, as in many organisms, natural selection leads to high levels of codon bias in genes that are highly expressed. Thus codon bias is an indicator of the intensity of one kind of selection that is experienced by genes and can be used to assess the impact of other genomic factors on natural selection. Among 13,000 genes in the Drosophila genome, codon bias has a slight positive, and strongly significant, association with recombination—as expected if recombination allows natural selection to act more efficiently when multiple linked sites segregate functional variation. The same reasoning leads to the expectation that the efficiency of selection, and thus average codon bias, should decline with gene density. However, this prediction is not confirmed. Levels of codon bias and gene expression are highest for those genes in an intermediate range of gene density, a pattern that may be the result of a tradeoff between the advantages for gene expression of close gene spacing and disadvantages arising from regulatory conflicts among tightly packed genes. These factors appear to overlay the more subtle effect of linkage among selected sites that gives rise to the association between recombination rate and codon bias.

THE redundancy of the genetic code has often been used to tease out manifestations of natural selection that would be beyond the resolution of most experimental approaches. The classic example is the contrast between substitution rates (and polymorphism levels) for mutations that do, and do not, alter the amino acid sequences of proteins. The latter class, which falls within the redundancy of the code, is commonly assumed to be selectively neutral and thus provides a baseline for interpretation of the tempo and mode of natural selection on those mutations that do alter proteins (Kimura 1977; Li *et al.* 1985; McDonald and Kreitman 1991; Whitfield *et al.* 1993).

Another application of genetic code redundancy to the study of natural selection relies on evidence that natural selection does indeed act on synonymous mutations (evidence that partly undermines methods that assume neutrality of synonymous mutations). The evidence is that genes that are expressed at high levels often show strongly biased codon usage in favor of those codons that correspond to the most common tRNAs (Chavancy *et al.* 1979; Ikemura 1985, 1991; Moriyama and Powell 1997). The association between codon bias and gene expression has long been apparent in single-celled organisms (Bennetzen and Hall 1982; Gouy and Gautier 1982; Sharp and Li 1987b; Sharp and Devine 1989) and has recently been shown to hold in several multicelled organisms, including *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (Duret and Mouchiroud 1999). The finding that, at least for some organisms, codon bias is partly due to natural selection to optimize gene expression means that codon bias can be used as an easily measured indicator of the intensity of natural selection. Furthermore, in those organisms where natural selection is a cause of high codon bias for some genes, the spectrum of codon bias values (which generally vary widely among the genes of a genome) offers a window into a critical range of selection intensities between levels where selection is ineffective (*i.e.*, not greater in impact than genetic drift and mutation) and levels where selection plays a large role in codon usage.

Here we employ this basic idea—that codon bias is an indicator of the overall impact of one kind of natural selection experienced by a gene (*i.e.*, selection for high gene expression)—to address how other aspects of the Drosophila genome interact with this type of natural selection. Of course not all genes need be highly expressed, and one reason that a gene may have low codon bias is simply that mutations that raise the expression level of that gene do not increase fitness. However, we can still inquire whether the codon bias distribution among genes varies as a function of genomic factors that may have an effect on gene expression or that may affect how well natural selection acts on gene expression.

Kliman and Hey (1993) used this approach to check the prediction that the effectiveness of natural selection is reduced by stochastic effects of natural selection on

[1]*Corresponding author:* Department of Genetics, Rutgers University, Nelson Biological Labs, 604 Allison Rd., Piscataway, NJ 08854-8082. E-mail: jhey@mbcl.rutgers.edu

linked genes (Hill and Robertson 1966; Felsenstein 1974). Net natural selection is expected to be higher on average, and stochastic effects due to linkage will be reduced, when recombination is acting. The effects of selection on linked sites can be individually large, as when a beneficial mutation sweeps to fixation, carrying with it a single haplotype (Maynard Smith and Haigh 1974), or individually more subtle, as when deleterious mutations remove linked sequences from the population (Charlesworth *et al.* 1993). But regardless of the sign or magnitude of selection coefficients, tight linkage should lead to conflicting selection pressures on linked sites that act in essentially stochastic fashion to retard the overall efficacy of natural selection. If this Hill-Robertson effect (Hill and Robertson 1966) is acting among the genes of Drosophila, and if selection is acting on codon usage, then codon bias should be lower, on average, among genes in genomic regions of low recombination. This prediction was borne out among a sample of 385 genes, 40 of which occurred in low recombining regions and had significantly lower mean codon bias (Kliman and Hey 1993). The same pattern also emerged in a later study of 537 genes (Comeron *et al.* 1999).

Note that this hypothesis does not specify what types of selection cause the Hill-Robertson effect, but rather that if it occurs, then the overall effectiveness of selection will be reduced. Any and all mutations with impacts on fitness and in negative linkage disequilibrium with one another will contribute to a Hill-Robertson effect. The method of Kliman and Hey (1993) and of Comeron *et al.* (1999) was to see if a Hill-Robertson effect reduces how well natural selection acts on mutations with slight effects on fitness (*i.e.*, synonymous mutations that affect codon bias and gene expression).

However, a significant confounding factor is that the preferred codons of Drosophila (*i.e.*, those that increase in frequency in highly biased and highly expressed genes) all end in either G or C, which in turn leads to a correlation between codon bias and GC content. Recently, Marais *et al.* (2001) reexamined the association between codon bias and recombination in Drosophila using the predicted genes of the genomic sequence and concluded that the apparent positive association was a by-product of positive covariation between each of these variables and GC content.

In recent years the concept that selection conflicts arise under tight linkage has played a central role in research on polymorphism levels in natural populations (Begun and Aquadro 1992; Charlesworth *et al.* 1993; McVean and Charlesworth 2000), as well as in research on the evolutionary origin and maintenance of recombination (Otto and Barton 1997; Hey 1998; Rice and Chippindale 2001). Here we revisit the question using the Drosophila genome sequence and taking particular care to assess the effects of both recombination and GC content on codon bias. We also examine a related prediction: If the Hill-Robertson effect reduces the effectiveness of selection (and thus mean codon bias) for genes in regions of low recombination, then it should also be acting in genes located in regions of high gene density. That is, we expect the Hill-Robertson effect to be apparent wherever the opportunity for recombination among multiple targets of selection is limited.

MATERIALS AND METHODS

Gene sequences: The nearly complete *D. melanogaster* genome sequence (Adams *et al.* 2000) reported in the 19 major GenBank files of Release 2 of the Berkeley Drosophila Genome Project was downloaded by FTP (http://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/Scaffolds/LARGE). These sequences constitute ~115 Mb and the large majority of Release 2 of the Drosophila genome sequence. The small files of Release 2 were not included because they provide little information on gene density. The header information in each GenBank file contains the locations of all amino acid coding sequences (CDS). A gene was excluded from the analysis if its CDS was incomplete or if, in rare cases, the length of its CDS was not a multiple of three. For 436 genes (3.3% of those used in our analyses), more than one CDS was given. In these cases we simply used the first CDS listed. To check whether this could affect the analyses we compared the codon bias, using the effective number of codons (ENC) measure (Wright 1990) of alternative CDSs for 336 alternatively spliced genes having two reported CDS values. No significant difference in codon bias was found in comparisons of the first and second CDS listed for each gene (mean ENC: first CDS = 47.72; second CDS = 47.90; $P = 0.40$; Kruskal-Wallis test). Occasionally, the CDSs of adjacent genes overlapped (usually on opposite strands). If three or more genes were reported to overlap one another, they were excluded from the analysis. In the end, 12,999 genes met the criteria for inclusion. Noncoding GC content (GC_{nc}) was determined from intron GC content and, for those genes without introns, from the immediately flanking noncoding DNA (1000 bp on either side—truncated as necessary if they extended into the coding region of an adjacent gene).

The data for all 12,999 genes were compiled into a common spreadsheet for statistical analysis. That spreadsheet is available from http://lifesci.rutgers.edu/~heylab.

Measuring codon bias: Codon bias in Drosophila and other organisms has been measured in two primary ways: departure from equal codon usage (Shields *et al.* 1988; Wright 1990; Akashi 1994; Sharp *et al.* 1995) and as a function of the frequencies of those codons that have been shown to be more common in genes that have more unequal codon usage (Sharp and Li 1987a; Stenico *et al.* 1994). Both types of measures are highly correlated, reflecting the presence of underlying causes that act in similar ways among the multiple codons of the different amino acids. In Drosophila, as in many microorganisms and some eukaryotes, a key underlying factor is gene expression, and it is this covariation that motivates a focus on codon bias as an indicator of the effectiveness of selection on a key phenotype. However, a second factor that can complicate many analyses is GC content. All of the preferred codons in Drosophila end in either G or C, and so codon bias indices covary with GC content in Drosophila (Shields *et al.* 1988; Akashi 1995). A third factor is gene length, which negatively covaries with codon bias in Drosophila (*i.e.*, longer genes have less codon bias; Powell and Moriyama 1997; Comeron *et al.* 1999).

As a measure of codon bias we used the primary factor found

in a factor analysis of relative codon usage. Factor analysis is a multivariate statistical technique, similar to principal components analysis, which is designed for cases when multiple variables (codon usage frequencies in the present case) are thought to be shaped by a common factor. To avoid a basic GC content factor and a gene-length factor, synonymous codon frequencies (calculated for each of the 59 codons that fall within a set of synonymous codons) were first regressed on GC_{nc} and on the length of the amino-acid-coding portions of genes, and the residuals from these regressions were used for the factor analysis. Factor analysis and the majority of other statistical analyses were done using SAS version 8.0.

Gene expression estimation: We used counts of reported expressed sequence tags (ESTs) as a rough indicator of the level of gene expression (Duret and Mouchiroud 1999), hereafter referred to as *E*. EST data were obtained from release 4.0 of the Drosophila Gene Index (DGI) of The Institute of Genome Research, available at http://www.tigr.org/tdb/dgi/. The DGI is a catalog of all multiple Drosophila EST data sets, organized by apparent sequence matching into tentative consensus (TC) sequences. The DGI also lists known or predicted gene names for TCs. These gene names were matched up with those in the Drosophila genome database. A TC was counted for a particular gene if the gene name identifier was found in the text string for that TC. Genes that did not appear among the TC identifiers were screened individually using gene name synonyms as given in Flybase (http://flybase.bio.indiana.edu/). For each gene, *E* is the mean of the EST counts among the TCs identified with that gene. Most genes corresponded to just a single TC, and the minority of genes with multiple TCs were primarily restricted to those listed as having alternative transcripts. TCs, and nonzero EST counts, were identified for 12,763 of the 12,999 genes in the study. *E* ranged from 0 to 834. Because the DGI includes ESTs from many different cDNA libraries, which vary in terms of represented life stages and tissues, *E* does not reflect the ways that genes may differ in timing and location of gene expression and should be considered only as a rough overall measure of gene expression.

Recombination estimation: Genome-wide recombination rate estimators (which generate estimates of the recombination rate per generation for every gene) can be generated from the relationship between genetic and physical maps of the Drosophila genome. To make sure that our findings were not an artifact of any particular approach, we considered five different measures of recombination.

KH93: This measure was developed by Kliman and Hey (1993) and is based on the derivative of four- and five-term polynomials of the relationship between the genetic and the physical maps for each chromosome. This measure was developed using genes with reported genetic map positions and polytene chromosome map positions in an early version of Flybase (Ashburner 1992). Actual physical distances between genes were based on optical density estimates of DNA content of polytene chromosomes (Sorsa 1988).

ACE: The adjusted coefficient of exchange is the measure used by C. Aquadro and colleagues (Begun and Aquadro 1992; Aquadro *et al.* 1994; Kindahl 1994) and is represented in Tables 1.2, 1.3, and 1.4 of Kindahl (1994). It consists of local estimates of the relationship between the genetic map and the physical map over numbers of polytene chromosome bands and, as for *KH93*, it relies upon the DNA content estimates of Sorsa (1988).

R_{TE}*:* This measure is based on genetic and physical map data collected by Ising and Block (1984) using the TE transposable element (Ising and Ramel 1976). These values have been used by the developers of Flybase to create a standard table of physical and genetic map locations (A. de Gray, personal communication). Because a common type of marker was used in a common Drosophila line, it was thought that these measures may be a more reliable basis for determining genetic map position, given a known physical location. The table is located at http://flybase.bio.indiana.edu/maps/lk/cytotable.txt. To use these measures, we also obtained the DNA base number of the cytological map positions from http://flybase.bio.indiana.edu:82/maps/lk/genome-cyto-seq-map/. We then plotted the genetic map against DNA position for each chromosome arm and for each cytological chromosome section and measured the slope of the genetic map over eight points flanking the cytological position. The estimate of slope was taken as the local recombination rate. Values for genes not in the data set were obtained by finding the closest gene that was listed in the table.

R: A current table of genes listing all those with independently determined genetic map locations was obtained from Flybase (A. de Gray, personal communication). Genes were excluded if they had multiple conflicting map estimates and if their genetic map position placed them out of sequence with nearby genes on the basis of the physical locations in the genome sequence. The resulting set of 493 X-linked and autosomal genes was ordered by physical position on each chromosome, and recombination was estimated by taking the slope, over eight flanking genes, of the genetic map position as a function of the DNA position on the chromosome. Recombination rates were then assigned to all the genes on the basis of those in the set of 493 to which they are closest.

R_P*:* This measure, like KH93, is based on four-term polynomial regressions of genetic map position on physical map location, with a separate regression done for each chromosome arm. However, R_P is based on the 493 loci used for *R*.

Gene density estimation: We focused primarily on two different estimates of gene density, each calculated from the base positions of genes indicated in the GenBank files of the genome sequence. "Space between genes" (SBG) is the mean of the distances on either side of a gene, that is, between a gene's terminal codon and the closest terminal codon of the nearest gene. "Genes per kilobase" (GPK) was calculated from the number of genes, including fractions of a gene, observed over a 20,000-base region centered on the midpoint of a gene. In addition, analyses were also conducted using other measures of gene density, including measures like GPK, but measured over longer and shorter intervals, as well as measures of codon density. All measures, including those based on codon density, behaved very similarly to those reported here (results available upon request).

RESULTS

Table 1 shows the correlation coefficients among all five measures of recombination. All are highly correlated with each other; all are similarly, weakly, but highly significantly, correlated with codon bias; and none are significantly correlated with GC_{nc}. Hereafter, analyses that included recombination are based on *R*, although all analyses using either R_P or R_{TE} are nearly identical to those obtained with *R* and those obtained using *KH93* and *ACE* are qualitatively very similar to those obtained with *R* (results available upon request).

The factor analysis among the regression residuals of the frequencies of 59 codons revealed a strong primary factor with an eigenvalue over three times that associated with the second factor. For 12,999 genes, the distri-

TABLE 1

Correlations among recombination estimates, GC_{nc}, and codon bias

	KH93	*ACE*	R_{TE}	*R*	R_P	GC_{nc}
ACE	0.698					
	$P < 0.0001$					
R_{TE}	0.7970	0.6997				
	$P < 0.0001$	$P < 0.0001$				
R	0.7873	0.7185	0.8253			
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$			
R_P	0.8783	0.7332	0.8768	0.8773		
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$		
GC_{nc}	−0.0077	−0.0084	−0.0092	−0.0056	−0.0034	
	$P = 0.3825$	$P = 0.3387$	$P = 0.293$	$P = 0.521$	$P = 0.701$	
F	0.0807	0.0500	0.0813	0.0829	0.0791	0.000
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	

Product-moment correlation coefficients and significance levels are shown among the recombination rate estimators, noncoding GC (GC_{nc}), and codon bias (*F*) described in MATERIALS AND METHODS. For all comparisons the sample size was 12,999.

bution of factor scores for this primary factor has a mean of zero and is closely approximated by a normal distribution (Figure 1). The factor scores were highly correlated with other measures of codon bias (Figure 1), and hereafter this primary factor, denoted as *F*, is used as our measure of codon bias. Because *F* is based on residuals from linear regression of codon frequencies against GC_{nc} and gene length, the product-moment correlation between *F* and GC_{nc}, as well as between *F* and gene length, is zero.

Figure 2 shows *F* plotted against the recombination rate measure *R* and gene expression *E* with means and 95% confidence intervals shown for each bin of ~1000 genes. Except for the lowest values of *E* (4383 genes had an EST count of either 0 or 1), gene expression is positively associated with *F*. The product-moment correlation between *F* and *E* is 0.146 ($P < 0.0001$).

From Figure 2 it appears that the association between *F* and *R* is primarily limited to the genes with the lowest values of recombination. The 4000 genes with the lowest estimated recombination rates have a product-moment correlation between *R* and *F* of 0.167 ($P < 0.0001$), whereas the correlation among the 9000 genes with the highest levels of *R* is −0.021 (not significant). The association between *F* and the lowest levels of recombination is especially apparent for the genes on the small fourth chromosome of Drosophila, which does not recombine. The fourth chromosome consists of interspersed regions of apparent euchromatin and heterochromatin (as evidenced by position effect variegation experienced by inserted genes; SUN *et al.* 2000). The 78 genes in the study from the fourth chromosome actually have a higher mean number of ESTs (10.7 ESTs/gene) than do the other genes (8.8 ESTs/gene), although this

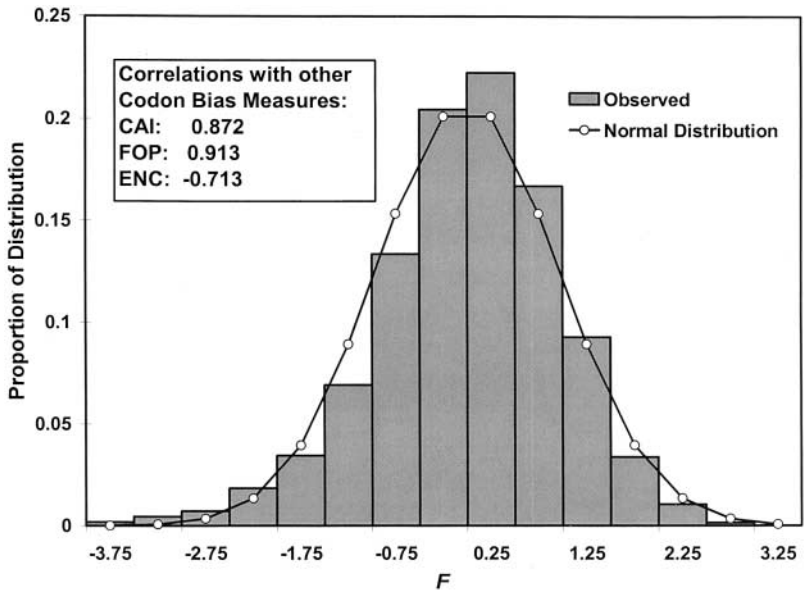


FIGURE 1.—The distribution of factor scores for *F*, the first primary factor of a factor analysis of 12,999 genes, and the expected proportions of a normal distribution with the same mean and variance. The mean is zero and the variance is 0.92. The product-moment correlation coefficient was calculated between factor scores and three other measures of codon bias: the ENC, which is high when codon bias is low (WRIGHT 1990), the CAI (SHARP and LI 1987a), and the FOP (SHARP and DEVINE 1989).

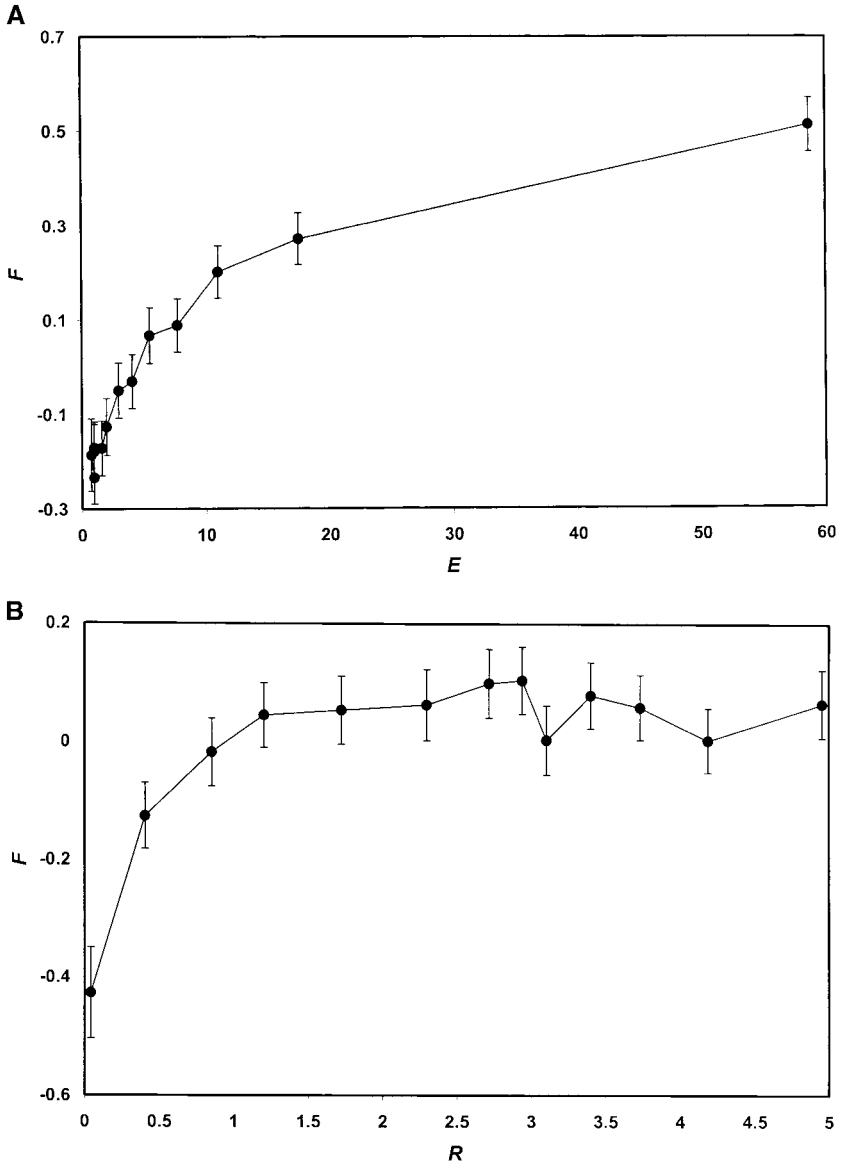


FIGURE 2.—(A) Codon bias (F) is plotted against a measure of gene expression (E). (B) Codon bias (F) is plotted against recombination rate (R). For each plot, genes were ordered by their values for the independent variables (E and R) and grouped into 13 bins, each of 1000 genes. Mean values and 95% confidence intervals for F were calculated for each bin and plotted against the corresponding means for the independent variable. The 95% confidence intervals are not shown with respect to the x-axis because these values generally span a very small distance (usually <2% of the position of the point on the x-axis).

is largely due to two genes with high EST counts and the difference in means is not statistically significant. Despite the absence of sharp differences in gene expression, the mean codon bias score was −2.32 for the fourth chromosome genes, a point far out in the left tail of the distribution (Figure 1) and a value far lower than the mean for the other genes (0.015, $P < 0.001$; Wilcoxon two-sample test). If these genes are removed from Figure 1, then the lower left point of the curve for R shifts upward to a mean value of F of −0.22.

The fourth chromosome genes are unusual in other respects, including a low mean value of GC_{nc} (0.312 *vs.* 0.377 for other genes; $P < 0.001$; Kruskal-Wallis test) and low values of gene density as measured by SBG (8998 *vs.* 5984 bp; $P < 0.001$; Kruskal-Wallis test) and GPK (0.092 genes/kb *vs.* 0.212 genes/kb; $P < 0.001$; Kruskal-Wallis test). However, the low level of codon bias on the fourth chromosome is not solely due to the reduced GC content. Considering only those 1000 nonfourth chromosome genes that have the lowest values of GC_{nc}, and that have a mean GC_{nc} less than that for the fourth chromosome loci, the mean value of F is 0.037, far higher than that for the fourth chromosome genes. In the majority of analyses to follow, the fourth chromosome genes are not included.

If conflicting selection pressures that arise from linkage disequilibrium among sites that are under selection actually lead to reduced codon bias because of a reduced efficiency of selection, then we would expect that genes that are physically closer to other genes would experience more selection conflicts and, thus, would also have lower codon bias. The reasoning is simply that since genes are the likely location of most mutations that have effects on fitness, then genes that are closest to each other should be those most likely to experience selection conflicts due to linkage. Figure 3A shows that codon bias varies as a complex curvilinear function of SBG, with a less curvilinear relationship with GPK. Surpris-

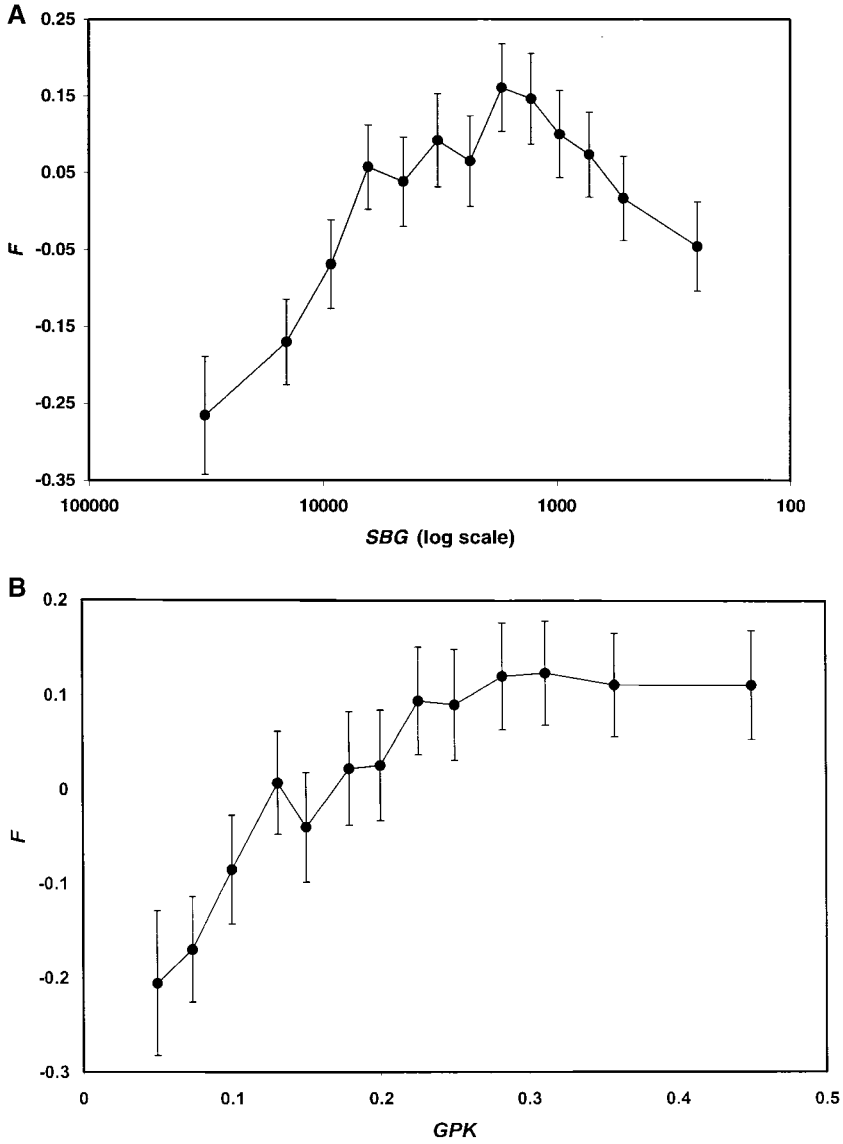


FIGURE 3.—Codon bias (F) is plotted against two measures of gene density: (A) SBG and (B) GPK. The 12,921 genes not located on the fourth chromosome were ordered, and mean values were calculated for 13 bins, each of ~1000 genes. Mean values and 95% confidence intervals for F are plotted against mean density values for each bin. In A and B, the x-axis orientation is set so that gene density increases to the right.

ingly, the major effect in both curves is an increase of codon bias with gene density, which is in the opposite direction of that expected by models of selection conflicts under linkage. Only for those genes for which there is little space separating them from flanking genes (low values of SBG) do we find the expected decline of codon bias with increasing gene density. As suggested by the 95% confidence intervals in Figure 3, the nonlinear relationships are highly significant. Quadratic polynomial regression of codon bias on the gene density measures reveals a highly significant curvilinear relationship. This is apparent for GPK ($F' = -0.401 + 3.14 \times \mathrm{GPK} - 4.54 \times \mathrm{GPK}^2$, where F' is the predicted value of codon bias) and for those genes with SBG values <5000 ($F' = -0.0467 + 0.000163 \times \mathrm{SBG} - 3.52 \times 10^{-9} \times \mathrm{SBG}^2$). Overall regression of F on SBG is complicated by the highly nonuniform distribution of SBG values. However, a quadratic polynomial regression of F on ranked values of SBG ($\mathrm{SBG_r}$) for all genes also reveals a downward curve ($F' = -0.357 + 1.23 \times 10^{-4} \times \mathrm{SBG_r} - 7.83 \times 10^{-9}\ \mathrm{SBG_r}^2$). In each of these cases, the fit of the quadratic model, as well as the value of each estimated regression parameter, is highly significant ($P < 0.001$).

Measures of gene spacing and density also covary with intron length and intron number in such a way that all these variables seem to reflect a common factor of gene packing. For genes that have introns, both total intron length and the number of introns are lower for genes that are more densely packed (Figure 4). For GPK some of this covariation is a necessary consequence of the variable itself (more and longer introns must reduce local gene density that is measured over a fixed distance), but the same pattern holds for the length of SBG.

The observation that more closely packed genes have higher codon bias is not due to covariation with noncoding GC content. $\mathrm{GC_{nc}}$ is positively correlated with SBG ($r = 0.0283$, $P = 0.0013$) and the correlation coefficient with GPK is negative, albeit not significantly different

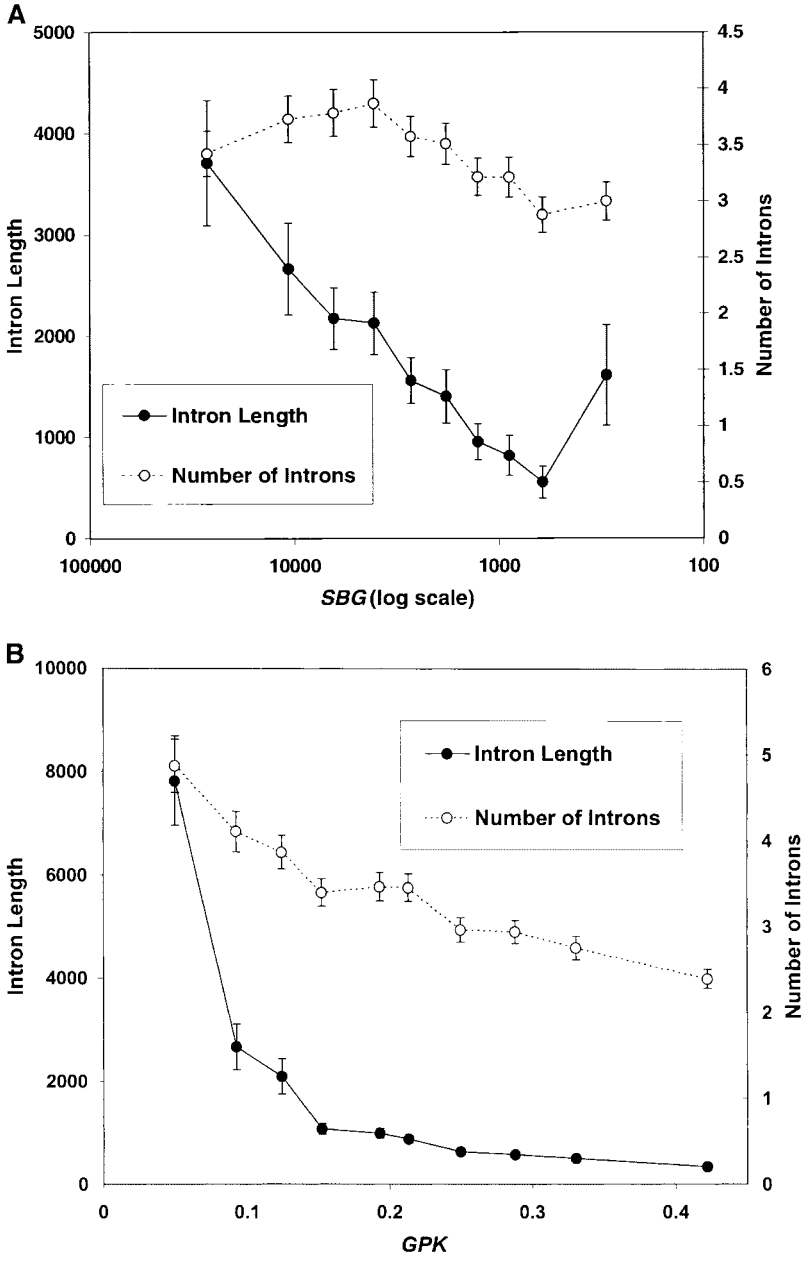


FIGURE 4.—Means and 95% confidence intervals for total intron length and the number of introns are plotted against mean values for (A) SBG and (B) GPK. Only those genes not on the fourth chromosome and that have introns are included.

from zero ($r = -0.010$, $P = 0.253$). Both of these correlations are in the opposite direction of that required to explain the observed covariation between gene density and codon bias. The predominant association between gene density and codon bias is probably not caused by selection conflicts associated with linkage, as the covariation between these variables is primarily in the wrong direction—higher gene density should lead to more selection conflicts and lower codon bias, the reverse of what is seen. To be sure, we checked to see if polymorphism levels were correlated with gene density, as they are with recombination rates (BEGUN and AQUADRO 1992; KINDAHL 1994). For polymorphism levels previously reported for 32 genes (BEGUN and AQUADRO 1992; KINDAHL 1994), the correlation coefficients do not approach statistical significance (for SBG, $r = -0.060$, $P = 0.708$, $n = 32$; for GPK, $r = -0.008$, $P = 0.967$, $n = 32$).

The relationship between codon bias and gene density is partly mirrored by the relationship between gene expression and measures of gene density. Figure 5 reveals a curvilinear relationship between E and ranked measures of gene density. The genes with the highest gene expression are, on average, those that occur at intermediate levels of gene density. As in the case of codon bias, the curvilinear relationships between gene expression and gene density are highly significant. Quadratic polynomial regression of E on the gene density measures reveals downward curvilinear relationships: $E' = 7.58 + 0.00286 \times \text{SBG} - 6.21 \times 10^{-7} \times \text{SBG}^2$ (for genes with SBG <5000 bp) and $E' = 4.86 + 34.4 \times \text{GPK} - 57.7 \times \text{GPK}^2$, where E' is the predicted value of

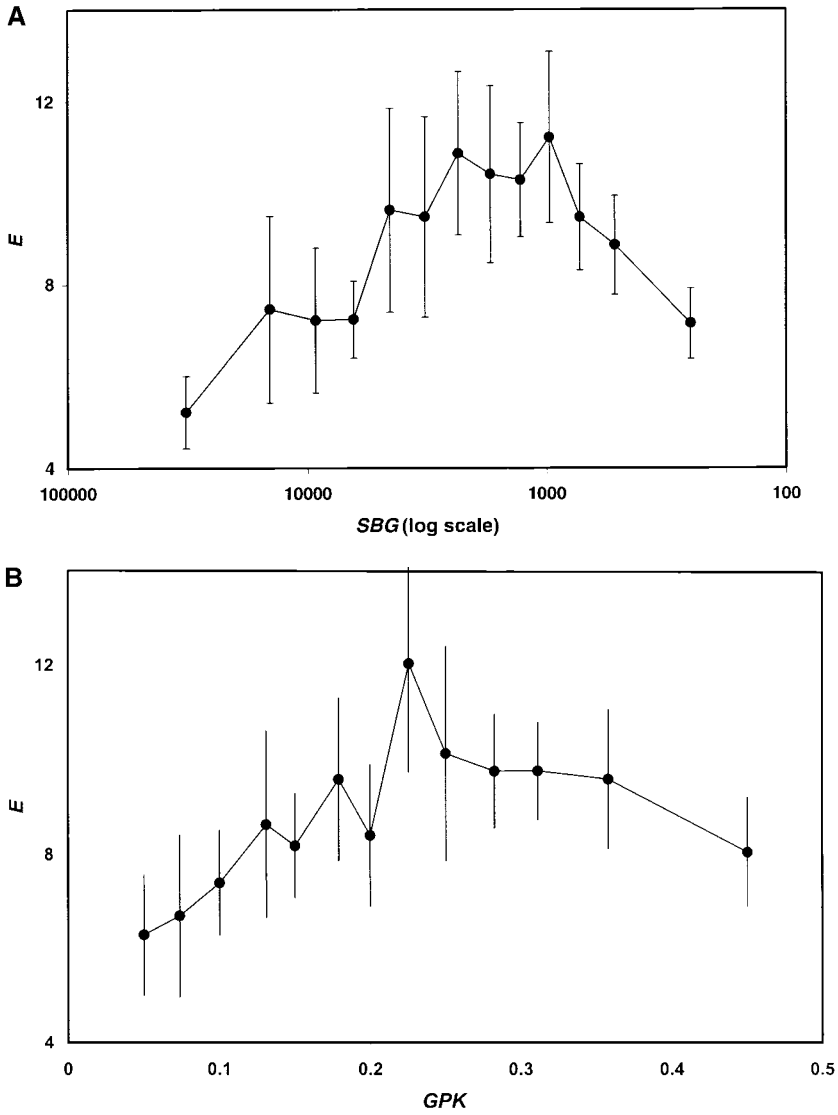


FIGURE 5.—Mean gene expression (E) and 95% confidence intervals for those genes not on the fourth chromosome are plotted against mean values of (A) SBG and (B) GPK.

gene expression. In each case, the fit of the quadratic model, as well as the value of each estimated regression parameter, is highly significant ($P < 0.001$). Interestingly, the genes with the very lowest values of SBG show a marked increase in intron length (Figure 4A). Part of the reason for this is that some genes contain other genes within their introns and thus have large introns and SBG values of zero.

DISCUSSION

In recent years the patterns of codon usage in the genes of Drosophila have often been used to study the interplay of natural selection and other factors, such as mutation and genetic drift, that shape the DNA sequences of genes (AKASHI 1995, 1997; KLIMAN 1999; BEGUN 2001). Net natural selection on any one gene will be elevated by the differential fitness contributions of different codon variants (*i.e.*, selection coefficients), yet in the case of synonymous codons, we can reasonably suppose that those coefficients span a range of very low values. For some genes, either the substitution of alternative synonymous codons will make no impact on gene function or such impact as occurs will in turn make no impact on fitness. These genes should reveal a pattern of codon usage that reflects the mutation process at synonymous base positions. For other genes, alternative synonymous codons do vary in their effects on gene function, and if these effects are sufficiently large and sufficiently consistent in direction, then natural selection will shape the distribution of codon usage.

Codon usage is but one factor that has an effect on gene function, and generally we expect it to be a very minor one relative to other gene components (*e.g.*, relative to the amino acid sequence and to the locations and sequences of regulatory sites). However, codon usage serves a unique role for investigators. Unlike the case with other aspects of a gene, investigators have a clear hypothesis about the pattern of codon usage and the magnitude of one kind of natural selection on a gene.

TABLE 2

Correlations among codon bias estimates and other variables

	F	ENC	FOP	CAI	GC_{nc}	Length
ENC	−0.713					
	$P < 0.0001$					
FOP	0.913	−0.768				
	$P < 0.0001$	$P < 0.0001$				
CAI	0.873	−0.811	0.932			
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$			
GC_{nc}	0.000	−0.258	0.252	0.222		
		$P < 0.0001$	$P < 0.0001$	$P < 0.0001$		
Length	0.000	0.190	−0.130	−0.146	−0.101	
		$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	
R	0.083	−0.075	0.059	0.070	−0.006	−0.001
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P = 0.521$	$P = 0.8936$

Product-moment correlation coefficients and significance levels are shown among codon bias measures (ENC, FOP, CAI, and *F*), noncoding GC (GC_{nc}), gene length, and one measure of recombination *R*. All measures are described in MATERIALS AND METHODS. For all comparisons the sample size was 12,999.

Codon usage can be determined precisely, given a DNA sequence, and specific predictions regarding the action of natural selection can be generated simply on the basis of the pattern of codon usage. In particular, the idea that some codons are "preferred" and that others are "unpreferred" has emerged as highly explanatory for the understanding of codon bias in general (SHARP and LI 1986, 1987a; AKASHI 1995), as well as of variation in codon bias among genes and genomes (KLIMAN and HEY 1993; COMERON *et al.* 1999), and for the understanding of substitution rates at synonymous sites within genes (AKASHI 1996; BEGUN 2001; MUNTE *et al.* 2001).

Codon bias and recombination: Codon bias (*F*) is highly significantly correlated with several measures of recombination (Table 1); however, the correlation coefficients are quite low, and the corresponding r^2 values are miniscule, indicating that recombination explains just a tiny fraction of the variation found among genes in codon bias. The degree of association is also not a constant one over the range of recombination, and genes that have the lowest levels of recombination show a much stronger relationship between codon bias and recombination. This pattern was also reported in previous studies with fewer genes (KLIMAN and HEY 1993; COMERON *et al.* 1999).

The low degree of association between recombination rate and codon bias means that the relationship can be easily obscured if those variables are measured with high variance or if there are confounding covariates. Consider the associations between other measures of codon bias and *R*, GC_{nc}, and gene length (Table 2). The ENC (WRIGHT 1990), the frequency of optimal codons (FOP; SHARP and DEVINE 1989), and the codon adaptation index (CAI; SHARP and LI 1987a) all reveal a weak correlation with *R*, as does *F*, yet all but *F* also reveal stronger correlations with GC_{nc} and gene length. The correlations in Table 2 also serve to make a different point: that it is possible for a measure of codon bias to be correlated with many factors. As long as each explains only a small portion of the variance in codon bias, as is clearly the case, then there remains ample scope for the other covariates.

The pattern of association between codon bias and recombination has important implications for models of the evolutionary origin and maintenance of recombination. If it is indeed the case that mutations that affect fitness are often in linkage disequilibrium simply due to the physical distance between them, then there arises a benefit to mutations that elevates the recombination rate (OTTO and BARTON 1997; HEY 1998). Also, if synonymous mutations do affect fitness, albeit weakly, then codon bias should serve as a sensitive indicator of persistent Hill-Robertson conflicts, if such effects do indeed occur. However, we find that most genes (the ~9000 genes with recombination rates >1.5 cM/Mb) show no association between recombination and codon bias. These genes appear to have recombination rates higher than necessary to dispel Hill-Robertson conflicts. In other words, selection conflicts that arise under disequilibrium due to linkage appear not to be a sufficient explanation of the high recombination rates found in most genes of Drosophila.

Recently MARAIS *et al.* (2001) examined the association between recombination and codon bias in Drosophila and concluded that, while there was a positive correlation, it was a by-product of common covariation with noncoding GC content. MARAIS *et al.* (2001) used FOP, which does have a strong association with GC_{nc}, as their measure of codon bias (Table 2). A key component of their conclusion was a finding that recombination rate is also correlated with noncoding GC content, a pattern that is not evident for any of the recombination measures considered here (Table 1). Fortunately, MARAIS *et al.* (2001) have made their primary spreadsheet

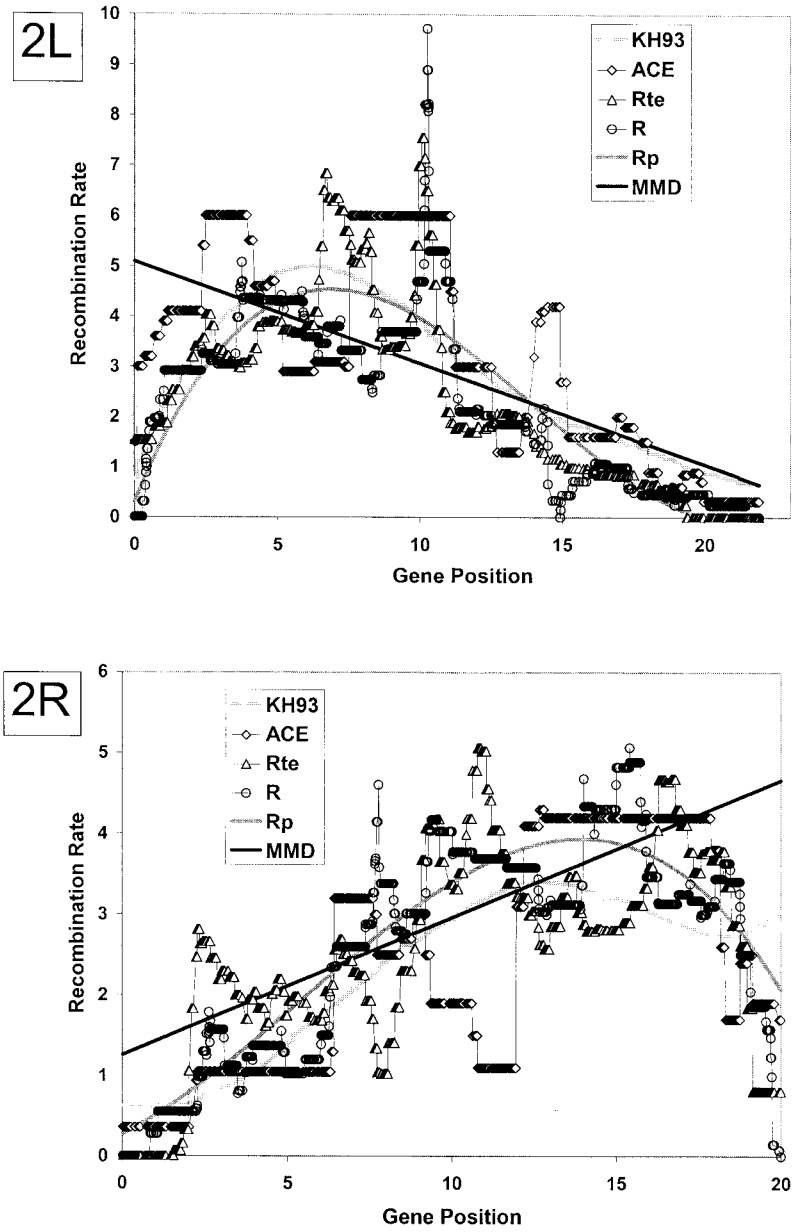
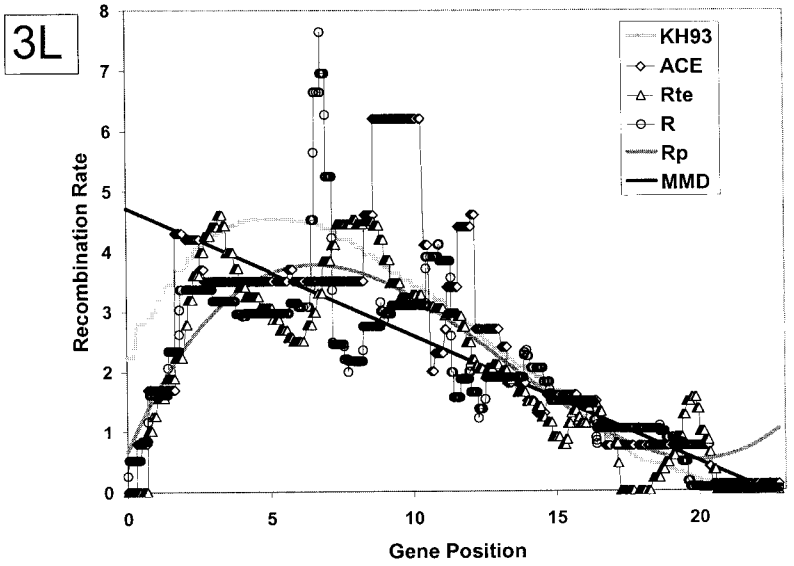
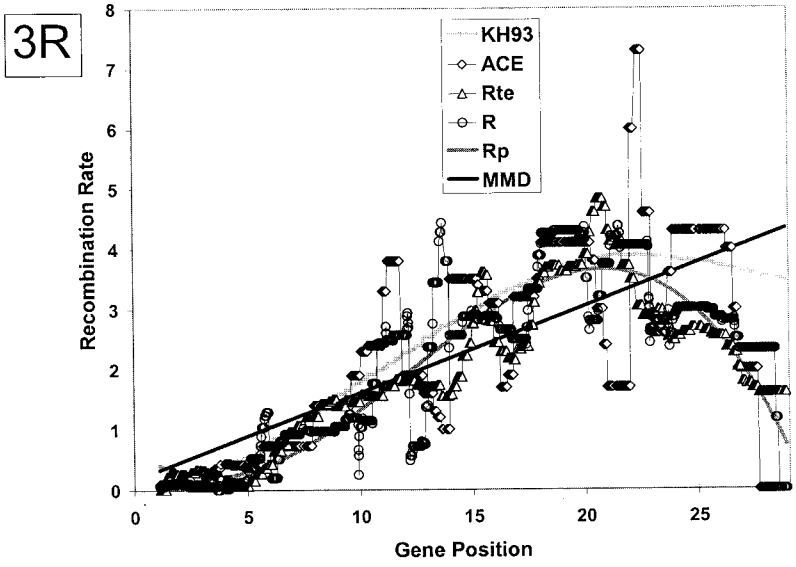


FIGURE 6.—Recombination rate estimates for those genes included in both the present study and that of MARAIS *et al.* (2001). For each chromosome arm (2L, 2R, 3R, 3L, and X) the recombination rate (centimorgans per megabase) is plotted against gene position (in megabases from the telomere). The recombination measures are those listed in MATERIALS AND METHODS, except for MMD, which is the measure used by MARAIS *et al.* (2001). To avoid clutter, those estimates based on polynomials are shown using lines, while those estimates based on short intervals are shown using symbols.

available, and it is a straightforward matter to compare their recombination estimates with those considered for the present study. The data of MARAIS *et al.* (2001) are from 13,877 genes and are available at http://pbil.univ-lyon1.fr/datasets/Marais2001/data_droso.txt. The two data sets were compiled separately and independently, and so for a variety of reasons not all genes have a match in both data sets. A total of 10,767 genes had matching names and lengths in both data sets.

The recombination estimates of MARAIS *et al.* (2001) are based on second-order polynomial regression of genetic maps on physical maps. Since the derivative of a second-order polynomial is necessarily the formula for a straight line, all five of their recombination functions (one per major chromosome arm) extend from a high telomeric value to a low centromeric value. Figure 6 shows, for each major chromosome arm (2L, 2R, 3L, 3R, and X), the recombination rate estimates considered for this article as well as those of MARAIS *et al.* (2001). Three patterns are apparent: (1) All of the recombination measures covary considerably for most portions of most chromosome arms; although (2) there is also considerable scatter among these measures, reflecting the approximate nature of these types of estimators; and (3) the linear estimator of MARAIS *et al.* (2001) departs markedly from the other estimators, particularly for the X chromosome and the telomeres of the other chromosomes. Most importantly, the central conclusion of MARAIS *et al.* (2001) that an association between GC content and recombination creates an apparent association between recombination and codon bias appears to be mistaken, because only their measure of recombina-





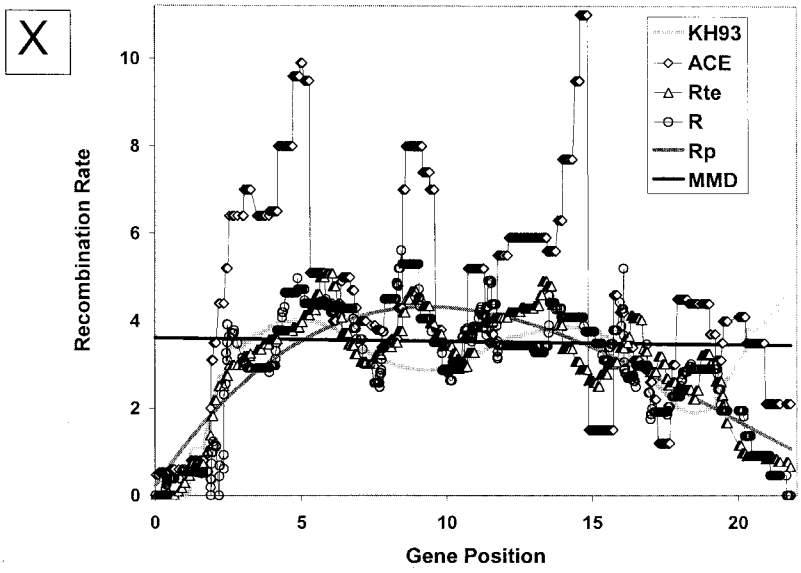


FIGURE 6.—*Continued.*

tion is correlated with noncoding GC content. This correlation arises by assigning high recombination rates to telomeric regions, which do indeed have high GC content, but which actually have reduced levels of recombination (Figure 4).

Marais *et al.* (2001) also argued that if codon bias covaried with recombination, independently of GC content, then we would find that the frequency of nonoptimal codons would decrease with recombination. They observed a negative correlation for *Fnop-AU* (the frequency of nonoptimal codons that end in A or U) and their measure of recombination, but a significant positive correlation ($r = 0.06$, $P < 0.0001$) between recombination and *Fnop-GC*. Using the argument that selection increases in efficiency with high recombination, and thus permits higher frequencies of optimal codons, and thus high codon bias, *Fnop-GC* should decrease with recombination, the reverse of what was found (Marais *et al.* 2001). We also measured *Fnop-GC* and *Fnop-AU* using the designations of optimal and nonoptimal on the basis of which codons increase or decrease in frequency along with gene expression, as given by Duret and Mouchiroud (1999). The correlation between R and *Fnop-AU* is negative, as expected ($r = -0.0393$, $P < 0.0001$) while the correlation between R and *Fnop-GC*, while not negative, is very near zero ($r = 0.0029$, $P = 0.7363$).

Codon bias, gene expression, and gene density: The prediction of an association between recombination and codon bias is based on the idea that genomic regions of low recombination will experience more linkage disequilibrium among mutations that have effects on fitness. If we assume that the regions within and near genes are where most such mutations occur, as opposed to regions between genes, then the same line of reasoning leads to the prediction that genes in regions of high gene density should also have more linkage disequilibrium between mutations with effects on fitness and should also reveal an association between codon bias and, in this case, gene density. However, this pattern was not observed, and we found instead a curvilinear relationship between measures of density (Figure 3) and recombination and between those measures and gene expression (Figure 5). As far as we have been able to determine from a review of the literature, these are novel observations. The presence of a curvilinear pattern suggests that multiple factors are acting to shape the relationship between F (and E) and measures of gene density. In what follows we develop a hypothesis of conflicting factors that predominate over different ranges of gene density.

In eukaryotes, including Drosophila, the local chromatin structure around genes is a major determinant of gene expression levels (Gasser *et al.* 1998; Farkas *et al.* 2000; Bell *et al.* 2001), and in *Saccharomyces cerevisiae* there is a tendency for adjacent and nearby genes to have similar expression patterns (Cohen *et al.* 2000). The results presented here on gene density are consistent with models in which local levels of chromatin structure and associated effects on gene expression extend over regions that affect multiple genes. The evidence of local position effects on gene expression may be related to position effect variegation, a phenomena generally and strongly identified with heterochromatin (Henikoff 1990). However, none of the patterns described here are mitigated by exclusion of genes near the centromeres (results available upon request). It should also be noted that regions of high and low gene density are interspersed widely across the X chromosome and the autosomes (Adams *et al.* 2000). If the opening up of chromatin, which is required for gene expression, tends to occur over regions that span multiple genes, then one way to facilitate gene expression is to have genes lie near other genes that are expressed at the same times, in the same tissues, and at similar levels. On average, highly expressed genes would be the ones that benefit most from this type of facilitation, simply because these genes are the ones most likely to have expression requirements that are partly coincident with those of other genes. According to this view we would expect that genes that are highly expressed would tend to lie near other genes and that this would introduce positive covariation between gene expression and density. This is what was observed over the lower half of the range of gene density (Figure 5).

This simple argument, if carried further, also entails a cost for high levels of gene packing. As genes move closer, they come closer to the regulatory domains of other genes, and their own regulatory elements (linked enhancers and silencers) become closer to other genes, which creates selection pressure for insulating elements that prevent regulatory conflicts (Bell *et al.* 2001). From Figure 5 it appears that the turnover point at which gene density maximally facilitates gene expression lies approximately at the eighth point from the left for both SBG and *GBK*. This corresponds to a mean distance between genes (SBG) of 1700 bp and a mean number of genes per kilobase (GPK) of 0.23.

This specific tradeoff model is a hypothesis motivated by the curvilinear patterns of Figure 3 and by literature reports on the large role played by chromatin in gene expression. However, given the simple circumstantial nature of the evidence, other hypotheses for all or part of the curvilinear pattern could certainly be developed. For example, transposable elements may tend to accumulate preferentially in regions where genes are not expressed at high levels and where recombination is low, and given their length, such insertions would tend to reduce gene density. It could also be the case that baseline rates of other types of insertions and deletions vary over the genome in such a way as to contribute to the observed patterns and that this is for reasons not associated with natural selection to optimize gene expression.

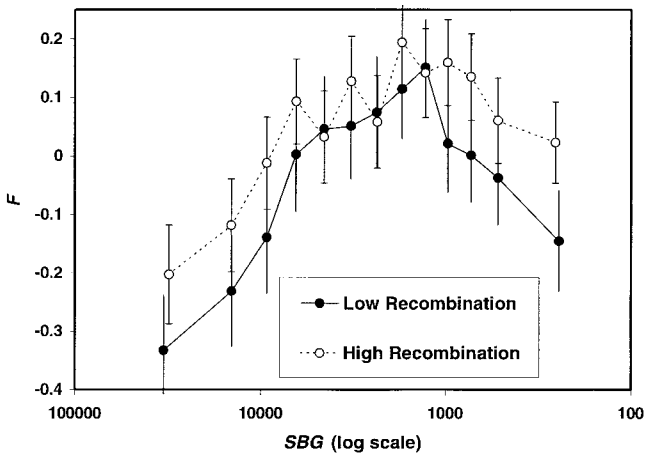


FIGURE 7.—Mean values and 95% confidence intervals for codon bias (F) are plotted against SBG for genes that have levels of recombination that are higher than the mean value of recombination ($R > 2.442$ cM/Mb) and for genes that have levels of recombination lower than the mean.

Hill-Robertson effects and gene density: In a complex multivariate context it can be difficult to tease out the pattern of interdependence among variables, particularly when patterns of covariation are not linear. In the present context we would especially like to understand whether Hill-Robertson effects and resulting selection pressures favoring recombination have played a role in shaping gene density. However, the complex patterns of covariation prevent clear conclusions. Nevertheless, three points can be made. First, the negative relationship between codon bias and gene density, found for the most tightly packed genes, is in the direction predicted by the Hill-Robertson effect. It is possible that the Hill-Robertson effect becomes strong only at the very highest levels of gene density, just as it appears to for the genes with lowest levels of recombination (Figure 2), and that this contributes to the positive association between gene density and both F and E for the highest levels of density. Second, if Hill-Robertson effects really do occur among genes with low recombination rates per base pair, then this will create a selection pressure to increase the space between genes. This point is essentially the same as that made to explain the observed negative correlation between intron length and recombination in Drosophila (CARVALHO and CLARK 1999; COMERON and KREITMAN 2000). If Hill-Robertson effects are strong among genes in regions of low recombination, then selection will favor longer introns in those genes (COMERON and KREITMAN 2000; DURET 2001) and more space between those genes. In fact, the correlation between R and SBG ($r = -0.043$, $P < 0.0001$) and between R and GPK ($r = 0.047$, $P < 0.001$) is small, but in the proper direction for such an effect. Of course this argument does not serve to explain why our indicator of selection efficacy (codon bias) increases with gene density over much of the density range.

The third point is that both gene density and recombination explain only small portions of the variance in codon bias and in gene expression and only small portions of the variance of each other. Thus it may be best to consider the effects of recombination on F and E as largely independent of the effects associated with gene density. To check this we assessed the correlation between F and R for low values of SBG (for SBG $\leq$1700 bp, $n = 5441$, $r = 0.0518$, $P < 0.0001$) and for high values of SBG (for SBG $>$1700 bp, $n = 7558$, $r = 0.1001$, $P < 0.0001$). These correlations changed negligibly when SBG was held constant via partial correlation (for SBG $\leq$1700 bp, $n = 5441$, $r = 0.0515$, $P < 0.0001$; for SBG $>$1700 bp, $n = 7558$, $r = 0.0963$, $P < 0.0001$). Finally, the independence of gene density and recombination is fairly striking in a simple plot of F against SBG for both high and low levels of recombination (Figure 7). The two curves closely parallel each other, with that for high recombination consistently higher than that for low recombination, and both curves have a peak in the middle of the span that is essentially identical in location to that for the combined data (Figure 3). These results are consistent with gene density and Hill-Robertson effects acting essentially independently on gene expression and codon bias. They also suggest that Hill-Robertson interference occurs across the gene density spectrum, as F is higher for high recombination genes across the spectrum of SBG values.

We are grateful to Hiroshi Akashi, James Birchler, Sarah Elgin, and Adam Eyre-Walker for helpful comments and to two reviewers for helpful critique and suggestions. This research was supported by National Institutes of Health grants R01GM54684 to J.H. and R15HG02456 to R.M.K.

LITERATURE CITED

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

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

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

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

AKASHI, H., 1997 Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205:** 269–278.

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination, and DNA polymorphism in Drosophila, pp. 46–56 in *Non-neutral Evolution*, edited by B. GOLDING. Chapman & Hall, London.

ASHBURNER, M., 1992 Flybase, a Drosophila genetic database, version 9209.

BEGUN, D. J., 2001 The frequency distribution of nucleotide variation in Drosophila simulans. Mol. Biol. Evol. **18:** 1343–1352.

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.

BELL, A. C., A. G. WEST and G. FELSENFELD, 2001 Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science **291:** 447–450.

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

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

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

Chavancy, G., A. Chevallier, A. Fournier and J. P. Garel, 1979 Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryotic cell. Biochimie **61:** 71–78.

Cohen, B. A., R. D. Mitra, J. D. Hughes and G. M. Church, 2000 A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. **26:** 183–186.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Duret, L., 2001 Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet. **17:** 172–175.

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Farkas, G., B. A. Leibovitch and S. C. Elgin, 2000 Chromatin organization and transcriptional control of gene expression in Drosophila. Gene **253:** 117–136.

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

Gasser, S. M., R. Paro, F. Stewart and R. Aasland, 1998 The genetics of epigenetics. Cell. Mol. Life Sci. **54:** 1–5.

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

Henikoff, S., 1990 Position-effect variegation after 60 years. Trends Genet. **6:** 422–426.

Hey, J., 1998 Selfish genes, pleiotropy and the origin of recombination. Genetics **149:** 2089–2097.

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

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

Ikemura, T., 1991 Correlation between codon usage and tRNA content in microorganisms, pp. 87–112 in *Transfer RNA in Protein Synthesis*, edited by D. L. Hatfield, B. J. Lee and R. M. Pirtle. CRC Press, Boca Raton, FL.

Ising, G., and K. Block, 1984 A transposon as a cytogenetic marker in *Drosophila melanogaster*. Mol. Gen. Genet. **196:** 6–16.

Ising, G., and C. Ramel, 1976 The behavior of a transposing element in *Drosophila melanogaster*, pp. 947–954 in *Genetics and Biology of the Drosophila*, edited by M. Ashburner and E. Novitski. Academic Press, London.

Kimura, M., 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267:** 275–276.

Kindahl, E. C., 1994 Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*. Ph.D. Thesis, Cornell University, Ithaca, NY.

Kliman, R. M., 1999 Recent selection on synonymous codon usage in Drosophila. J. Mol. Evol. **49:** 343–351.

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

Li, W. H., C. I. Wu and C. C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol**. 2:** 150–174.

Marais, G., D. Mouchiroud and L. Duret, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

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

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

Moriyama, E. N., and J. R. Powell, 1997 Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45:** 514–523.

Munte, A., M. Aguadé and C. Segarra, 2001 Changes in the recombinational environment affect divergence in the yellow gene of Drosophila. Mol. Biol. Evol. **18:** 1045–1056.

Otto, S. P., and N. H. Barton, 1997 The evolution of recombination: removing the limits to natural selection. Genetics **147:** 879–906.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Rice, W. R., and A. K. Chippindale, 2001 Sexual recombination and the power of natural selection. Science **294:** 555–559.

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, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

Sharp, P. M., and W. H. Li, 1987a The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1294.

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

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi and J. F. Peden, 1995 DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **349:** 241–247.

Shields, D., 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.

Sorsa, V., 1988 *Chromosome Maps of Drosophila*. CRC Press, Boca Raton, FL.

Stenico, M., A. T. Lloyd and P. M. Sharp, 1994 Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res. **22:** 2437–2446.

Sun, F. L., M. H. Cuaycong, C. A. Craig, L. L. Wallrath, J. Locke *et al.*, 2000 The fourth chromosome of *Drosophila melanogaster*: interspersed euchromatic and heterochromatic domains. Proc. Natl. Acad. Sci. USA **97:** 5340–5345.

Whitfield, L. S., R. Lovell-Badge and P. N. Goodfellow, 1993 Rapid sequence evolution of the mammalian sex-determining gene sry. Nature **364:** 713–715.

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

Communicating editor: S. W. Schaeffer