

# Intragenic Hill-Robertson Interference Influences Selection Intensity on Synonymous Mutations in *Drosophila*

Josep M. Comeron and Theodore B. Guthrie

Department of Biological Sciences, University of Iowa

Natural selection influences synonymous mutations and synonymous codon usage in many eukaryotes to improve the efficiency of translation in highly expressed genes. Recent studies of gene composition in eukaryotes have shown that codon usage also varies independently of expression levels, both among genes and at the intragenic level. Here, we investigate rates of evolution ( $K_s$ ) and intensity of selection ( $\gamma_s$ ) on synonymous mutations in two groups of genes that differ greatly in the length of their exons, but with equivalent levels of gene expression and rates of crossing-over in *Drosophila melanogaster*. We estimate  $\gamma_s$  using patterns of divergence and polymorphism in 50 *Drosophila* genes (100 kb of coding sequence) to take into account possible variation in mutation trends across the genome, among genes or among codons. We show that genes with long exons exhibit higher  $K_s$  and reduced  $\gamma_s$  compared to genes with short exons. We also show that  $K_s$  and  $\gamma_s$  vary significantly across long exons, with higher  $K_s$  and reduced  $\gamma_s$  in the central region compared to flanking regions of the same exons, hence indicating that the difference between genes with short and long exons can be mostly attributed to the central region of these long exons. Although amino acid composition can also play a significant role when estimating  $K_s$  and  $\gamma_s$ , our analyses show that the differences in  $K_s$  and  $\gamma_s$  between genes with short and long exons and across long exons cannot be explained by differences in protein composition. All these results are consistent with the Interference Selection (IS) model that proposes that the Hill-Robertson (HR) effect caused by many weakly selected mutations has detectable evolutionary consequences at the intragenic level in genomes with recombination. Under the IS model, exon size and exon-intron structure influence the effectiveness of selection, with long exons showing reduced effectiveness of selection when compared to small exons and the central region of long exons showing reduced intensity of selection compared to flanking coding regions. Finally, our results further stress the need to consider selection on synonymous mutations and its variation—among and across genes and exons—in studies of protein evolution.

## Introduction

Synonymous codons are not used randomly in many species, including several model eukaryotes such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and humans (Ikemura 1985; Sharp and Li 1987; Kliman and Hey 1993; Moriyama and Hartl 1993; Akashi 1994; Akashi 1995; Akashi and Schaeffer 1997; Powell and Moriyama 1997; Comeron, Kreitman, and Aguade 1999; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000; Duret 2000; Akashi 2003; Kliman, Irving, and Santiago 2003; Comeron 2004; Plotkin, Robins, and Levine 2004). In all these species, highly transcribed genes exhibit a more biased usage of synonymous codons, favoring codons that correspond to abundant tRNAs. These observations suggest that natural selection influences synonymous codon usage to improve the efficiency of translation in highly expressed genes (i.e., translational selection), by increasing accuracy and/or speed of translation (Sharp and Li 1986; Bulmer 1991; Akashi 1994; Carlini and Stephan 2003). Selection on synonymous mutations can also act at the level of mRNA folding stability (Bulmer 1991; Duan et al. 2003), with strong secondary structures avoided in highly transcribed genes (Carlini, Chen, and Stephan 2001). Variation in levels of gene expression is commonly proposed as the main selective cause for differences in codon bias among genes.

Recent analyses of synonymous base composition have suggested additional sources of variation in codon bias in eukaryotes. First, there is an association between the length of the coding sequence (CDS) and synonymous

codon usage that cannot be explained only by differences in expression or location in the genome (Moriyama and Powell 1998; Comeron, Kreitman, and Aguade 1999; Duret and Mouchiroud 1999; Comeron and Kreitman 2002; Comeron 2004). Second, intragenic analyses in *D. melanogaster* reveal variation in codon usage along exons (Kliman and Eyre-Walker 1998; Comeron and Kreitman 2002; Qin et al. 2004). In particular, codon bias is reduced in the central region of long exons, a U-shaped distribution along the CDS that is not observed in genes with introns (Comeron and Kreitman 2002). Third, all else being equal, genes with different exon-intron structure show different intragenic patterns and overall degrees of bias in codon usage (Comeron and Kreitman 2002; Qin et al. 2004). Altogether, these observations suggest that there are mutational or selective factors influencing synonymous composition beyond those associated with translational selection or other features that might vary across genomes. Such factors ought to be associated, directly or indirectly, with exon size and the exon-intron structure of genes.

A possible selective explanation for these observations is associated with the evolutionary consequences of selection on linked sites—i. e., the Hill-Robertson (HR) effect (Hill and Robertson 1966; Felsenstein 1974). The HR effect establishes that selection acting on one site will reduce the effectiveness of selection on linked sites in finite populations. The consequences of HR can be investigated—as an acceptable simplification—as being equivalent to a reduction in the effective population size,  $N_e$ , that ultimately reduces the efficacy of selection (Hill and Robertson 1966; Felsenstein 1974; Charlesworth, Morgan, and Charlesworth 1993; Kliman and Hey 1993; Kondrashov 1994; Barton 1995; Caballero and Santiago 1995; Otto and Barton 1997; Hey and Kliman 2002). Significantly, many sites under weak selection—if genetically close to each other—might cause

Key words: codon bias, weak selection, gene structure, gene conversion bias, interference selection.

Email: josep-comeron@uiowa.edu

*Mol. Biol. Evol.* 22(12):2519–2530. 2005

doi:10.1093/molbev/msi246

Advance Access publication August 24, 2005

a detectable HR effect, reducing  $N_e$  and the effectiveness of selection (Li 1987; Comeron, Kreitman, and Aguade 1999; McVean and Charlesworth 2000; Tachida 2000; Weinreich and Rand 2000; Piganeau et al. 2001; Betancourt and Pre-graves 2002; Comeron and Kreitman 2002). We use the term Interference Selection (IS) to specify the HR effect caused by many weakly selected mutations, which has distinct evolutionary properties compared to HR caused by strongly selected mutations (see McVean and Charlesworth 2000, Comeron and Kreitman 2002).

In most genomes, sites under weak or moderate selection are likely to be physically clustered, mostly in coding and in regulatory regions. On the other hand, the evolutionary consequences of IS caused by clusters of sites under weak selection are predicted to spread over very short genetic distances in genomes with recombination. In most eukaryotes, therefore, the causes and consequences of IS will be local, at the level of individual exons or genes, and can generate heterogeneity in the effectiveness of selection among genes in association with exon sizes and exon-intron structures as well as across long exons (Comeron and Kreitman 2002).

Understanding selection on synonymous mutations is significant because it sheds light on the evolutionary forces operating on a great number of extant mutations in many species (i.e., in the hundreds of thousands or millions in most eukaryotes). It has also significant consequences for many evolutionary studies because synonymous mutations are often used to gauge the neutral mutation rate when screening for fingerprints of natural selection (Nielsen and Yang 1998; Yang et al. 2000; Bustamante et al. 2002; Fay, Wyckoff, and Wu 2002; Smith and Eyre-Walker 2002).

Here, we investigated two specific predictions of the IS model: (1) long exons will show reduced effectiveness of selection when compared to small exons, and (2) the central region of long exons will show reduced intensity of selection compared to flanking coding regions of the same exons. To this purpose, we obtained and analyzed divergence and polymorphism data in *Drosophila* species, focusing on synonymous mutations because their population and evolutionary features are expected to be very sensitive to small changes in  $N_e$ . Specifically, we estimated and compared rates of evolution ( $K_s$ ) and the intensity of selection ( $\gamma_s$ ) on synonymous mutations in two groups of genes that differ greatly in the length of their exons. We also studied the possible heterogeneity in  $K_s$  and  $\gamma_s$  across long exons. To estimate  $\gamma_s$ , we applied two methods that are independent of mutation rates and patterns: the first based on the ratio of polymorphism to fixed divergence ( $rpd$ ; Sawyer and Hartl, 1992), the second based on polymorphic data only to obtain more contemporary estimates of  $\gamma_s$  (Maside, Lee, and Charlesworth 2004).

## Materials and Methods

### *Drosophila* Lines, Genes, DNA Isolation, and Sequencing

Fifty genes were sequenced in six *Drosophila* lines: one *D. melanogaster* (Ral-1, kindly provided by T. C. MacKay), two African *Drosophila simulans* (kindly provided by W. Ballard), two *Drosophila yakuba* (Tai18 and

**Table 1**  
**List of Genes Investigated<sup>a</sup>**

Locus	Length (bp) of CDS in <i>Drosophila melanogaster</i>	Locus	Length (bp) of CDS in <i>Drosophila melanogaster</i>
CG10321	2,508	CG14842	534
CG14411	2,532	CG17304	534
CG3615	2,538	CG7960	534
CG11770	2,577	CG13221	537
CG14514	2,619	CG31624	537
CG12283	2,643	CG9617	537
CG7093	2,646	CG6030	537
CG9211	2,661	CG7197	540
CG4977	2,685	CG8365	540
CG13350	2,688	CG15696	540
CG18471	2,703	CG10839	549
CG2899	2,901	CG8343	549
CG17075	2,907	CG1738	549
CG5669	2,907	CG13308	549
CG6407	3,015	CG4764	549
CG10719	3,114	CG5012	549
CG32217	3,180	CG2911	552
CG12234	3,726	CG2257	552
CG18265	3,969	CG4185	552
CG6890	4,041	CG1956	555
CG8896	4,158	CG14745	555
CG12105	4,227	CG14746	558
CG11156	4,278	CG13448	558
CG8595	4,341	CG8577	558
CG17766	4,578		

<sup>a</sup> Genes with a single translated exon.

T33; kindly provided by A. Llopart) and one *Drosophila erecta* (Tucson *Drosophila* Stock Center, stock number 14021-0224.0). DNA was isolated following standard procedures (Ashburner 1989). A single DNA fragment encompassing the complete CDS was amplified by polymerase chain reaction (PCR) for each gene and cleaned up using Wizard MagneSil PCR System (Promega, Madison, Wisc.). We sequenced directly both strands of each PCR product using Applied Biosystems Big Dye Terminator (version 3.0/3.1) chemistry on ABI PRISM 3100/3730 Genetic Analyzers (Applied Biosystems, Foster City, Calif.). Gene definitions and sequences for the second *D. melanogaster* line were obtained from the *Drosophila* genome (<http://www.flybase.org>, Version 2.0/3.0) (Adams et al. 2000). Only genes supported by empirical data (e.g., full-length mRNA information) and with no evidence of multiply spliced forms were investigated. Table 1 specifies the 50 genes studied. All newly obtained DNA sequences are deposited in EMBL/GenBank data libraries (accession numbers: DQ138644–DQ138943).

### Divergence Estimates

We estimated the number of synonymous ( $K_s$ ) and nonsynonymous ( $K_a$ ) substitutions per site using two different methods. First, we used an approximate method that imposes no assumption on the selective/neutral effect of mutations to estimate  $K_s$  and  $K_a$  and is fairly robust to deviations from equal base composition (Li 1993; Comeron 1995). These approximate estimates of  $K_s$  and  $K_a$  use the ratio of transitions to transversions ( $r$ ) as a single mutational parameter (Kimura 1981). The two sets of genes under scrutiny in this study exhibit similar but not identical values of

$r$  (1.93 and 2.06 for genes with short and long CDS, respectively). Variation in  $r$  can influence estimates of  $Ks$  and  $Ka$  because  $r$  has an effect on the estimated number of synonymous and nonsynonymous sites; higher  $r$  will increase the number of estimated synonymous sites and therefore underestimate  $Ks$ . If the difference in  $r$  plays a role on our estimates of  $Ks$ , the observation that genes with long CDS have higher  $Ks$  (see *Results and Discussion*) is then conservative because they also show higher  $r$ . The method applied to correct for multiple hits at a site is unlikely to play a role because there is no extreme bias in nucleotide composition, and the most distant comparison shows only  $Ks \sim 0.23$  (after correction). Finally, the influence that selection on nonsynonymous mutations may have on estimates of  $Ks$  is only quantifiable when many codons differ in two or more sites between species (only 1% of codons in our study). Thus, differences in  $Ks$  are unlikely to be caused by the approximate method we applied. We obtained confidence values for  $Ks$  and  $Ka$  after 10,000 independent replicates using K-Estimator v6 (Comeron 1999), which takes into account the number of sites under study, amino acid composition,  $r$ , the number of estimated substitutions (Poisson distributed), and the variance generated by multiple hits at a site.

The second method used to investigate  $Ks$  ( $Ks_{ML}$ ) and  $Ka$  ( $Ka_{ML}$ ) is the maximum likelihood approach implemented in PAML 3.13 (Goldman and Yang 1994; Yang 1997). The influence (if any) that selection on synonymous or nonsynonymous mutations has on estimates of  $Ks$  using this ML approach has not been investigated. This approach also requires information on mutation patterns between nucleotides to estimate the number of synonymous and nonsynonymous sites and therefore it can generate different estimates of  $Ks$  (even when very closely related sequences are compared) as a consequence of variation in the estimated number of synonymous sites. To have comparable  $Ks_{ML}$  and  $Ka_{ML}$  between two groups of genes (e.g., genes with short exons and genes with long exons) we applied the same equilibrium codon frequencies to both groups from the overall data set and fixed ratio transition-transversion ( $r = 2$ ), using the F3x4 model of codeml.

Overall  $Ks$  and  $Ka$  estimates were obtained using concatenated sequences of genes with short and long CDS. The total  $Ks$  and  $Ka$  along the whole phylogenetic reconstruction for *D. melanogaster*–*D. simulans*–*D. yakuba*–*D. erecta* species was obtained using the average distance in species with more than one sequence.

### Estimates of Selection on Synonymous Mutations

Under the infinitely many sites model, diploidy, and genic selection (Kimura 1964; Crow and Kimura 1970; Li 1987), the expected number of polymorphisms detected when investigating  $n$  chromosomes is

$$Pol(n, \gamma) = \theta \int_0^1 (1 - x^n - (1 - x)^n) \frac{1 - e^{-\gamma(1-x)}}{(1 - e^{-\gamma})x(1-x)} dx,$$

where  $\gamma$  is a measure of the selection coefficient ( $s$ ) scaled by the diploid effective population size ( $N_e$ ;  $\gamma = 4 N_e s$ ) and  $\theta = 4 N_e u$ , with  $u$  indicating the mutation rate per generation.

The expected number of fixed differences observed between two species that diverged  $t_{div}$  generations ago is

$$\begin{aligned} Fixed(t_{div}, n_1, n_2, \gamma) &= 2t_{div}u \frac{\gamma}{1 - e^{-\gamma}} \\ &+ \theta_1 \int_0^1 x^{n_1} \frac{1 - e^{-\gamma(1-x)}}{(1 - e^{-\gamma})x(1-x)} dx \\ &+ \theta_2 \int_0^1 x^{n_2} \frac{1 - e^{-\gamma(1-x)}}{(1 - e^{-\gamma})x(1-x)} dx, \end{aligned}$$

where  $n_1$  and  $n_2$  indicate the number of chromosomes investigated in species 1 and 2, respectively.

Following Sawyer and Hartl (1992), the ratio of polymorphism to fixed divergence ( $rpd$ ) allows us to estimate  $\gamma$  independently of the mutation rates or patterns

$$rpd = \frac{Fn}{\frac{T_{divN}}{4} + Gn},$$

using for simplicity  $Fn$  and  $Gn$  for  $\sum_{i=1}^j (1/\gamma)F(i, n_i, \gamma)$  and  $\sum_{i=1}^j (1/\gamma)G(i, n_i, \gamma)$ , respectively, with

$$F(i, n_i, \gamma) = \int_0^1 (1 - x^{n_i} - (1 - x)^{n_i}) \frac{1 - e^{-\gamma(1-x)}}{x(1-x)} dx,$$

$$G(i, n_i, \gamma) = \int_0^1 x^{n_i-1} \frac{1 - e^{-\gamma(1-x)}}{(1-x)} dx,$$

where  $j$  ( $j \geq 2$ ) indicates the number of species under study with polymorphism data,  $n_i$  the sample size in species  $i$  ( $n_i \geq 1$ ), and  $T_{divN}$  the time of the whole phylogenetic relationship for the species under study expressed as the number of  $N_e$  generations (Sawyer and Hartl 1992; Bustamante et al. 2002). Unbiased estimates of  $\gamma$ , however, require appropriate measures of  $t_{div}$  or  $T_{divN}$ , which can only be obtained by studying sites evolving neutrally (Sawyer and Hartl 1992). For instance, using mutations under weak selection to estimate  $t_{div}$  or  $T_{divN}$  tends to overestimate  $\gamma$  on favored mutations and underestimate  $\gamma$  on deleterious mutations.

When investigating synonymous changes, we can take advantage of the notion that synonymous codons can be classified as favored or nonfavored according to the influence of gene expression in their relative presence in a gene. Codons that increase (decrease) their presence with expression are presumed to cause beneficial (deleterious) effects on fitness. We can characterize the selection coefficient associated with a change from nonfavored to favored codon (preferred mutation, P) and from favored to nonfavored codon (unpreferred mutation, U) with the same magnitude and opposite sign. Under genic selection, the relative fitness of genotypes PP, PU, and UU is assumed to be  $1 - s$ ,  $1$ , and  $1 + s$ , respectively. Therefore, we can obtain the expectations of  $rpd$  for P ( $rpd_P$ ) and U ( $rpd_U$ ) mutations

$$rpd_P = \frac{Fn_P}{\frac{T_{divN}}{4} + Gn_P} \text{ and } rpd_U = \frac{Fn_U}{\frac{T_{divN}}{4} + Gn_U},$$

where the sign of  $\gamma$  is negative for U mutations ( $Gn_U$  and  $Fn_U$ ). Because  $T_{divN}$  is a common parameter for  $rpdp$  and  $rpdu$ , we obtain

$$\frac{Fn_P}{rpdp} - Gn_P = \frac{Fn_U}{rpdu} - Gn_U, \quad (1)$$

an expression that can be solved numerically to estimate the only variable,  $\gamma$ , if we estimate  $rpdp$  and  $rpdu$  from the data. We will use  $\gamma_s$  to indicate the magnitude of the selection intensity associated with P or U mutations and  $\gamma_{s-rpd}$  to indicate  $\gamma_s$  obtained from  $rpd$  data.

The second method to estimate  $\gamma_s$  is based on the relative presence of U and P polymorphic derived mutations (Maside, Lee, and Charlesworth 2004). This approach is also independent of mutation rates and patterns and allows us to obtain contemporary estimates of  $\gamma_s$ . Let  $P$  be the fraction of favored codons in a sequence where  $u_1$  and  $u_2$  are mutation rates at favored and nonfavored sites that will generate U and P mutations, respectively

$$P = \frac{e^\gamma}{k + e^\gamma},$$

with  $k = u_1/u_2$  (Li 1987; Bulmer 1991; McVean and Charlesworth 1999). The ratio of polymorphic U mutations to polymorphic P mutations ( $f$ ) observed when studying  $n$  chromosomes is

$$f = \frac{Pu_1 Pol(n, -\gamma)}{(1-P)u_2 Pol(n, \gamma)}.$$

In general terms  $f$  observed when studying  $j$  species, with  $n_i$  indicating the sample size in species  $i$ , is simply

$$f = \frac{-Fn_U}{Fn_P},$$

using for simplicity  $Fn_P$  and  $Fn_U$  for  $\sum_{i=1}^j F(i, n_i, \gamma)$  and  $\sum_{i=1}^j F(i, n_i, -\gamma)$ , respectively. We will use  $\gamma_{s-f}$  to indicate  $\gamma_s$  obtained from polymorphism data only. Note that  $\gamma_{s-f}$  increases with  $f$ , with  $f = 1$  when  $\gamma_s \rightarrow 0$ .

Here, we investigated polymorphism in three different species based on only two sequences each. Clearly, the study of a large number of sequences in several species would be desirable because it would increase the statistical power to detect possible differences in selection intensity based on both the number of polymorphic variants and the frequency of these variants. Nevertheless, our more limited approach has some features worthy of note. First, several of the most widely applied tests to detect the action of natural selection, including  $rpd$ , use the observed number of polymorphisms and not the frequency of the variants, and, all else being equal, the number of polymorphisms detected by analyzing a given number of chromosomes in a single species is expected to be smaller than if the total number of chromosomes is split into several species. Second, the analysis of a small number of sequences in several species allows a better inference of derived and ancestral states as well as the distinction between fixed and polymorphic variants, distinctions that are fundamental in  $rpd$  analyses. Third, it is one way to account, in part, for specific population history that could distort studies based on polymorphism in only one species.

## Confidence Intervals for $\gamma_s$

We applied the two proposed approaches to obtain confidence intervals for  $\gamma_s$ . First, for  $\gamma_{s-rpd}$  we obtained confidence limits under the Poisson random field (PRF) model. This model assumes that each site under study evolves independently, that each population has reached mutation-selection-drift (MSD) equilibrium, and that each of the four parameters (polymorphism and fixed divergence for P and U mutations) is an independent Poisson random variable (Sawyer and Hartl 1992; Bustamante et al. 2002). We obtained the posterior distribution for  $\gamma_{s-rpd}$  based on the Markov Chain Monte-Carlo (MCMC) method using the MKPRF program after 10,000 iterations (1,000 burn-in iterations, <http://cbruapps.tc.cornell.edu/mkprf.aspx>) (Bustamante et al. 2002; Barrier et al. 2003; Bustamante, Nielsen, and Hartl 2003). Following Maside, Lee and Charlesworth (2004), we obtained confidence intervals for  $\gamma_{s-f}$  also under the assumption of independence between sites and MSD equilibrium, based on binomial sampling of  $f$ .

## Favored and Nonfavored Codons

We defined favored and nonfavored codons in *D. melanogaster* following Duret and Mouchiroud (1999). This particular set of codons seems particularly appropriate because  $rpd$  estimates for changes presumed to be neutral (N changes), between two nonfavored codons, are the same for genes with short and long CDS (see *Results and Discussion*). Two other slightly different sets of favored and nonfavored codons proposed for *D. melanogaster* (Akashi 1994) and *D. simulans* (Akashi and Schaeffer 1997) cause  $rpdu$  to somewhat vary among sets of genes, in disagreement with the assumption of complete neutrality and therefore were not used in the analyses. Synonymous mutations with putative fitness effects (P and U mutations) were only studied when the ancestral and derived states could be resolved. Note that we assumed that codon preferences are constant for all of the species under study, within the *melanogaster* subgroup. This is in agreement with the observed conservation of codon preferences between *D. melanogaster* and *D. yakuba* as well as between *D. melanogaster* and more distant species of the *obscura* group (Kreitman and Antezana 2000; Powell et al. 2003).

## Results and Discussion

We obtained divergence and polymorphism sequence data in two sets of genes that differ in the length of the CDS but maintain the same exon-intron structure (i.e., genes with only one translated exon). We focused on 24 genes with short CDS (CDS between 530–560 bp) and 26 genes with long CDS (CDS longer than 2,500 bp). This second set of genes was also used to study intragenic patterns. We sequenced a total of 584 kb of CDS in six *Drosophila* lines (one line in *D. melanogaster*, two lines in *D. simulans*, two lines in *D. yakuba*, and one line in *D. erecta*); we included in the analyses the *D. melanogaster* complete nuclear sequence (FLYBASE 1998; <http://flybase.bio.indiana.edu/>). After alignment of the seven sequences, we analyzed 85,134 bp for genes with long CDS and 12,204 bp for genes with short CDS. The two sets of genes under

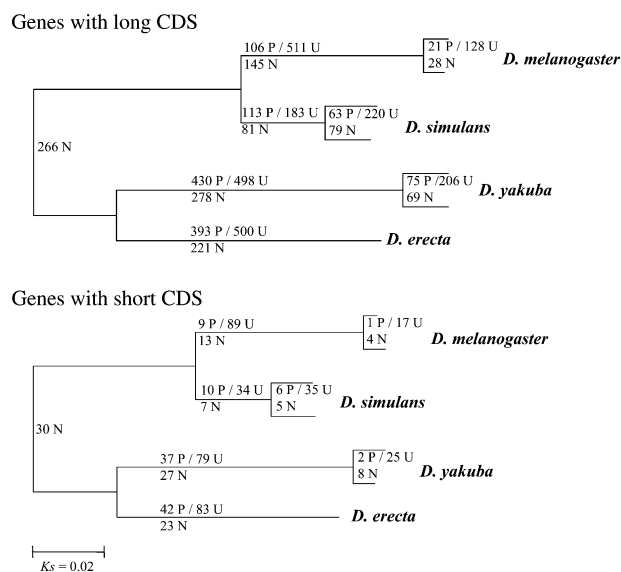


FIG. 1.—Schematic representation of the phylogenetic relationship between the sequences under analysis for genes with short and long exons, separately. Phylogenetic reconstruction obtained with neighbor-joining distance at synonymous sites. The total number of synonymous P, U, and N mutations unambiguously assigned to each lineage are also indicated.

study share similar rates of crossing-over in *D. melanogaster* (Mann-Whitney  $U$  test,  $Z = 0.41$ ,  $P > 0.65$ ). The comparison of transcription levels, using microarray expression data in *D. melanogaster* (Stolc et al. 2004), indicates no statistical difference between genes with short and long CDS (Mann-Whitney  $U$  test,  $Z = 0.78$ ,  $P > 0.40$ ).

#### Codon Usage and Synonymous Evolution

In agreement with the observations of Akashi (1996), the analysis of all 50 genes reveals that the bias in the use of synonymous codons (codon bias) is reduced in *D. melanogaster* (55.2% favored codons) when compared to *D. simulans* (56.6% favored codons). We also find that both *D. yakuba* (56.6% favored codons) and *D. erecta* (56.8% favored codons) show a pattern equivalent to that of *D. simulans*, underlining the idea that the pattern observed in *D. melanogaster* is derived.

The analysis of all 50 genes reveals 6,337 synonymous mutations that we classified according to their expected effects on fitness (Akashi 1995; Duret and Mouchiroud 1999). Unpreferred (U) mutations are deleterious changes from a favored to a nonfavored codon while preferred (P) mutations are selectively advantageous changes from a nonfavored to a favored codon. Mutations between two nonfavored codons, N, are assumed to be neutral to selection. Figure 1 shows a schematic representation of the phylogenetic relationship between the species under analysis and the number of P, U, and N mutations in each lineage. In total, we detected 4,169 synonymous mutations that can be classified as P or U; 3,177 fixed between species (1,140 P and 1,977 U) and 992 polymorphic in one species (168 P and 631 U). In agreement with the presumed deleterious effects of U mutations compared to P mutations, the

ratio of polymorphism to divergence ( $rp_d$ ) for U mutations is much higher than that for P mutations ( $rp_{dU} = 0.32$  and  $rp_{dP} = 0.15$ ). A McDonald-Kreitman test (McDonald and Kreitman 1991) for polymorphic and fixed P and U mutations indicates a significant departure from neutral expectations ( $G = 73.6$ ,  $P < 1 \times 10^{-10}$ ), confirming the selective nature of many synonymous mutations in *Drosophila*.

#### Comparison of Genes with Short and Long CDS

Genes with short CDS show a higher percentage of favored synonymous codons than genes with long CDS (63.9% and 55.1%, respectively;  $G$ -test of independence,  $P = 1 \times 10^{-12}$ ). This result is in agreement with earlier reports in *Drosophila* showing a negative relationship between CDS length and measures of codon bias (see *Introduction*). We estimated rates of synonymous ( $K_s$ ) and nonsynonymous ( $K_a$ ) evolution for each species pair and for the whole (unrooted) phylogenetic reconstruction (Table 2). The analysis of the whole phylogeny indicates that genes with short CDS show a significantly smaller  $K_s$  than that of genes with long CDS:  $K_{s(\text{short CDS})} = 0.288$  and  $K_{s(\text{long CDS})} = 0.346$  ( $P < 0.0001$ ; see *Materials and Methods*). This trend is not caused by individual lineages, as it is observed in most pairwise species comparisons, with the only exception being the *D. melanogaster*–*D. simulans* pair for which the difference in  $K_s$  is only barely significant ( $P = 0.018$ ). On the other hand, we find no difference in  $K_a$  ( $K_{a(\text{short CDS})} = 0.031$  and  $K_{a(\text{long CDS})} = 0.029$ ), which indirectly also suggests that the observed high  $K_s$  in genes with long CDS is not a consequence of undetected introns or inaccurate gene annotation. Equivalent results and conclusions are obtained for  $K_{sML}$  and  $K_{aML}$  using the maximum likelihood approach implemented in PAML 3.13 (Goldman and Yang 1994; Yang 1997) (see Table 2).

The observed difference in  $K_s$  between genes with short and long CDS is another indication of variation in selection intensities on synonymous mutations among genes. However, variation in mutation patterns associated with the length of the CDS and/or amino acid composition could also affect rates of evolution (McVean and Charlesworth 1999; Takano-Shimizu 2001). We, therefore, can take advantage of the fact that ratio of polymorphism to divergence ( $rp_d$ ) takes into account possible mutational differences across the genome, among genes or among codons to investigate variation in selection intensities (Sawyer and Hartl 1992; Akashi 1995; Akashi 1999; Bustamante et al. 2002).  $rp_d$  for N mutations ( $rp_{dN}$ ) is equivalent for genes with short and long CDS (0.170 and 0.177, respectively), as expected if most N mutations are neutral.

We estimated selection intensity ( $\gamma_s$ ) for P and U mutations based on the analytical predictions of  $rp_d$  (see *Materials and Methods* for details). Our methodology differs from that used in previous studies that inferred selection on synonymous mutations in that we use both types of mutations (P and U) simultaneously to estimate  $\gamma_s$ , an approach that eliminates the need to include estimates of the  $t_{div}$  or  $t_{divN}$ . Our  $rp_d$  analysis of the two classes of genes shows  $\gamma_{s-rp_d(\text{short CDS})} = 1.84$  and  $\gamma_{s-rp_d(\text{long CDS})} = 1.36$ ; 35% stronger selection intensity in genes with short CDS than in those with long CDS. This result is not due to the presence in

**Table 2**  
**Synonymous (Ks) and Nonsynonymous (Ka) Substitutions Per Site in Genes with Short and Long CDS<sup>a</sup>**

		Short CDS		Long CDS	
		Ks <sup>b</sup>	Ka	Ks	Ka
<i>Drosophila melanogaster</i> – <i>Drosophila simulans</i>	App.	0.085 (*)	0.009 (NS)	0.096	0.009
	ML	0.114 (NS)	0.008 (NS)	0.113	0.009
<i>Drosophila melanogaster</i> – <i>Drosophila yakuba</i>	App.	0.193 (***)	0.017 (NS)	0.233	0.019
	ML	0.241 (***)	0.016 (NS)	0.279	0.018
<i>Drosophila melanogaster</i> – <i>Drosophila erecta</i>	App.	0.182 (***)	0.019 (NS)	0.216	0.018
	ML	0.230 (**)	0.018 (NS)	0.260	0.017
<i>Drosophila simulans</i> – <i>Drosophila yakuba</i>	App.	0.175 (***)	0.018 (NS)	0.213	0.017
	ML	0.222 (***)	0.016 (NS)	0.256	0.017
<i>Drosophila simulans</i> – <i>Drosophila erecta</i>	App.	0.162 (***)	0.019 (NS)	0.193	0.017
	ML	0.207 (**)	0.018 (NS)	0.233	0.016
<i>Drosophila yakuba</i> – <i>Drosophila erecta</i>	App.	0.135 (***)	0.016 (NS)	0.169	0.014
	ML	0.176 (**)	0.014 (NS)	0.202	0.014
Total divergence	App.	0.288 (***)	0.031(NS)	0.346	0.029
	95% Confidence interval	0.267–0.310	0.027–0.034	0.337–0.355	0.028–0.031
	ML	0.370 (**)	0.028 (NS)	0.415	0.028
	95% Confidence interval	0.340–0.400	0.025–0.032	0.405–0.426	0.027–0.030

<sup>a</sup> App. and ML refer to estimates obtained using the approximate and maximum likelihood methods, respectively (see *Materials and Methods*).

<sup>b</sup> The probability that Ks or Ka in genes with short CDS is significantly different than that estimated in genes with long CDS after 10,000 simulations (see *Materials and Methods*) is given in parentheses.

\*  $P < 0.05$ , \*\*  $P < 0.01$ , \*\*\*  $P < 0.001$ , NS, nonsignificant,  $P > 0.05$ .

our sample of a few genes with extreme differences in  $\gamma_{s-rpd}$  because a nonparametric test based on ranks of order such as the Mann-Whitney  $U$  test rejects the assumption that  $\gamma_{s-rpd}$  from each of the genes with short CDS and those from each of the genes with long CDS come from the same sample ( $Z = 5.01$ ,  $P = 4 \times 10^{-8}$ ). The study of confidence limits for overall estimates of  $\gamma_{s-rpd}$  using a MCMC method under the PRF model (Sawyer and Hartl 1992; Bustamante et al. 2002; Barrier et al. 2003; Bustamante, Nielsen, and Hartl 2003) also reveals that  $\gamma_{s-rpd}$  in long genes is significantly smaller than that observed in short genes ( $P < 0.0001$ ; Table 3). Incidentally, MCMC results also show that  $\gamma_{s-rpd}$  is significantly greater than 0 in both classes of genes ( $P < 0.0001$  in both cases).

We then estimated more contemporary values for  $\gamma_s$  based on the ratio of U to P polymorphic mutations ( $f$ ; see *Materials and Methods* for details). The study of all polymorphic P and U mutations shows  $\gamma_{s-f(\text{short CDS})} = 3.1$  and  $\gamma_{s-f(\text{long CDS})} = 1.85$ , again indicating greater  $\gamma_s$  in genes with short CDS. A Mann-Whitney  $U$  test using  $\gamma_{s-f}$  for each gene individually rejects the possibility that this result is the consequence of a small number of genes with extreme behavior ( $Z = 4.4$ ,  $P = 3 \times 10^{-6}$ ). The probability of  $\gamma_{s-f(\text{short CDS})}$  being greater than  $\gamma_{s-f(\text{long CDS})}$  is 0.0011 (Table 3). In all, these results suggest a

selective, not mutational, cause to the observed difference between genes that differ in the length of their CDS.

#### Selection Across Long Exons

We studied three nonoverlapping coding regions in genes with long CDS: the 150 most proximal (5') codons, the 150 central codons, and the 150 most terminal (3') codons. In our set of genes we observe that the central region shows a significantly reduced codon bias (51.7% of favored codons) compared to both 5' (56.9%) and 3' (58.7%) flanking regions ( $G$ -test of independence;  $P < 1 \times 10^{-5}$  and  $P < 1 \times 10^{-8}$ , respectively). We also observe an increased Ks in the central region compared to both flanking regions, with Ks = 0.304, 0.380, and 0.327 for the 5', central, and 3' regions, respectively. Ks in the central region is significantly greater than that observed in the 5' or 3' regions of these long exons ( $P < 0.0001$  in both cases). There is no significant difference between the Ka values of central (0.032) versus flanking regions (0.030 and 0.031, for the 5' and 3' regions, respectively).

We then considered variation in  $\gamma_s$  across long exons, noting that any heterogeneity in  $\gamma_s$  cannot be attributed to differences in rates of transcription or mutational tendencies that vary across genomes. Estimates of  $\gamma_{s-rpd}$  for the three

**Table 3**  
**Estimates of Selection Intensity on Synonymous Mutations ( $\gamma_s$ )**

	Short CDS $P [\gamma_s = 0]$	Long CDS $P [\gamma_s = 0]$	Overall $\gamma_s$ $P [\gamma_{s-\text{long CDS}} \geq \gamma_{s-\text{short CDS}}]$
$\gamma_{s-rpd}$ (m/s/y/e) <sup>a</sup>	1.84 ( $P < 0.0001$ )	1.36 ( $P < 0.0001$ )	$P < 0.0001$
$\gamma_{s-rpd}$ (s/y/e)	2.06 ( $P = 0.0003$ )	1.58 ( $P < 0.0001$ )	$P < 0.0001$
$\gamma_{s-f}$ (m/s/y)	3.1 ( $P < 1 \times 10^{-10}$ )	1.85 ( $P < 1 \times 10^{-10}$ )	$P = 0.0011$
$\gamma_{s-f}$ (s/y)	2.94 ( $P < 1 \times 10^{-10}$ )	1.68 ( $P < 1 \times 10^{-10}$ )	$P = 0.0032$

<sup>a</sup> Estimates of  $\gamma_s$  using data from the species indicated in parentheses (m, *Drosophila melanogaster*; s, *Drosophila simulans*; y, *Drosophila yakuba*; e, *Drosophila erecta*). Probabilities for  $\gamma_{s-rpd}$  obtained under the PRF model (Bustamante et al. 2002) and probabilities for  $\gamma_{s-f}$  obtained following Maside, Lee and Charlesworth (2004).

**Table 4**  
**Relationship Between the Length of the CDS and the Relative Presence of Amino Acids Based on the Complete *Drosophila melanogaster* Genome**

Amino acid	Pearson's <i>R</i>	<i>P</i> Value
Asn	+0.100	$<1 \times 10^{-12}$
Asp	+0.080	$4 \times 10^{-11}$
Cys	-0.048	$3 \times 10^{-5}$
Gln	+0.150	$<1 \times 10^{-12}$
Glu	+0.107	$<1 \times 10^{-12}$
His	+0.111	$<1 \times 10^{-12}$
Lys	-0.018	NS
Phe	-0.013	NS
Tyr	-0.019	NS
Ile	-0.012	NS
Ala	+0.016	NS
Gly	-0.029	NS
Pro	+0.095	$<1 \times 10^{-12}$
Thr	+0.105	$<1 \times 10^{-12}$
Val	-0.037	0.001
Arg	-0.039	0.0007
Leu	+0.087	$<1 \times 10^{-12}$
Ser	+0.158	$<1 \times 10^{-12}$
Met	-0.047	$4 \times 10^{-5}$
Trp	-0.034	NS

NOTE.—NS, nonsignificant ( $P > 0.05$ ).

nonoverlapping regions reveal reduced selection in the central region, with estimates of 2.41, 1.42, and 1.82 for the 5', central, and 3' regions of long exons, respectively. Confidence intervals of  $\gamma_{s-rpd}$  based on MCMC (see *Materials and Methods*) indicate that  $\gamma_{s-rpd}$  in the central region is significantly smaller than that in the 5' ( $P = 0.0002$ ) or 3' ( $P = 0.0087$ ) flanking regions. Estimates of  $\gamma_s$  using polymorphic mutations also show reduced selection intensity in the central region, with  $\gamma_{s-f}$  of 2.92, 1.74, and 2.64 for the 5', central, and 3' regions of long exons, respectively.  $\gamma_{s-f}$  in the central region is significantly smaller than  $\gamma_{s-f}$  in either flanking region ( $P = 0.0037$  and  $P = 0.023$  when comparing to 5' and 3' flanking regions, respectively) or when compared to both flanking regions combined ( $P = 0.0004$ ). Furthermore, the central region of long exons shows a significantly reduced  $\gamma_{s-f}$  compared with that estimated for genes with short CDS ( $P = 0.0016$ ). In contrast, both flanking regions in long exons show estimates of  $\gamma_{s-f}$  similar to that estimated for genes with short CDS ( $P > 0.20$ ). Thus, the difference in  $\gamma_s$  observed between genes with short and long exons can be attributed, to a considerable degree, to the reduced  $\gamma_s$  observed in the central region of long exons.

#### Influence of Amino Acid Composition on Synonymous Evolution

Amino acid composition also plays a role in the observed variation in  $K_s$  and  $\gamma_s$  among and within genes. In fact, McVean and Vieira (2001) showed considerable variation among amino acids in the strength of selection acting on synonymous mutations using divergence data in *Drosophila* (McVean and Vieira 2001). We investigated whether amino acid composition could be associated with the difference in  $\gamma_s$  observed in our previous analyses. We first considered the possibility that genes with different

**Table 5**  
**Synonymous Substitutions Per Site ( $K_s$ )<sup>a</sup> for Different Amino Acids**

Amino acid	$K_s$ Genes with Short CDS	$K_s$ Genes with Long CDS
Ala	0.292	0.304
Arg	0.269	0.360
Asn	0.205	0.201
Asp	0.187	0.206
Cys	0.136	0.169
Gln	0.201	0.145
Glu	0.188	0.170
Gly	0.289	0.359
His	0.112	0.172
Ile	0.455	0.425
Leu	0.168	0.230
Lys	0.127	0.204
Phe	0.249	0.318
Pro	0.294	0.346
Ser2	0.126	0.185
Ser4	0.252	0.321
Thr	0.259	0.298
Tyr	0.133	0.187
Val	0.262	0.298
Twofold	0.171	0.185
Fourfold	0.252	0.306
Sixfold	0.252	0.343

<sup>a</sup>  $K_s$  based on the whole (unrooted) phylogenetic reconstruction for *Drosophila melanogaster*–*Drosophila simulans*–*Drosophila yakuba*–*Drosophila erecta*.

lengths of CDS have different amino acid composition. Our data set of 50 genes shows a significant difference in amino acid composition between genes with short and long CDS ( $P < 1 \times 10^{-10}$ ), as well as within long exons ( $P < 1 \times 10^{-5}$  comparing central and peripheral coding regions). This trend is also observed when the complete genome of *D. melanogaster* is analyzed; the relative presence of 11 out of the 20 amino acids is strongly associated with the length of the CDS (Table 4). Thus, amino acid composition does vary systematically with the length of the CDS and within genes, potentially influencing our interpretation when observing differences in  $K_s$  and  $\gamma_s$ .

We therefore estimated  $K_s$  for each amino acid separately (Table 5). The data show that  $K_s$  values vary among amino acids more than twofold, both in genes with short and long CDS. Interestingly, most amino acids show faster synonymous evolution when present in a long CDS than when present in a shorter CDS (Wilcoxon matched-pairs test,  $P = 0.003$ ). This effect is observed in twofold, fourfold, and sixfold degenerate amino acids. On average, the same amino acid shows 21% higher  $K_s$  when present in a long CDS than if present in a short CDS in our data set. To take into account possible mutational trends associated with context-dependent bias (Chen et al. 2004), we then estimated  $\gamma_{s-rpd}$  for each amino acid separately. As shown in Figure 2, different amino acids have different  $\gamma_{s-rpd}$ , particularly in genes with long CDS. When estimates of  $\gamma_{s-rpd}$  for a given amino acid are compared between genes with short and long CDS, amino acids in genes with short CDS show greater  $\gamma_{s-rpd}$  in 17 out of the 18 comparisons ( $P = 0.00087$ ), with the notable exception of cysteine. (Estimates of  $\gamma_{s-f}$  for different amino acids in genes with short CDS are not possible due to many instances with zero P polymorphic mutations.)

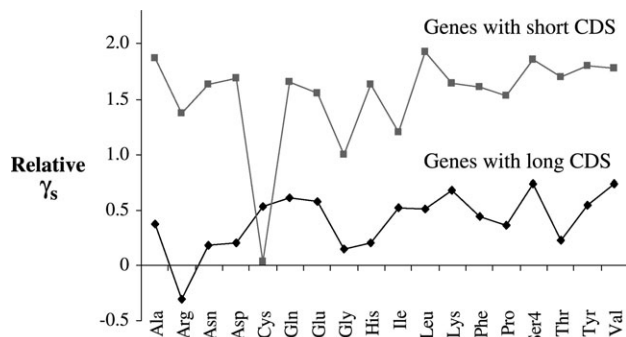


FIG. 2.—Estimates of selection intensity on synonymous mutations ( $\gamma_{s-rpd}$ ) for different amino acids and for genes with short and long CDS, relative to that of glycine.

In all, our study shows that amino acid composition does have an important effect on estimates of  $K_s$  and  $\gamma_s$  when comparing genes or gene classes. The observed difference between genes with short and long CDS, however, cannot be attributed to differences in amino acid composition and indicates actual differences in selection intensity.

#### Selective Causes for Heterogeneity in $\gamma_s$

Our results on  $K_s$  and  $\gamma_s$  are consistent with the two main predictions of the IS model: (1) the effectiveness of selection will decrease with exon size, and (2) the effectiveness of selection will show a U-shaped distribution along long exons (Comeron and Kreitman 2002). These results are also in complete agreement with previous intragenic studies of codon bias in *Drosophila*. We note however that a similar outcome could also result from relaxed selective constraints at the protein level in the central region of any long transcript, where the longer the transcript, the stronger the reduction in selective constraints on internal amino acid sites. If selection on synonymous mutations acts at the level of translational accuracy, this possibility also predicts a U-shaped distribution of  $\gamma_s$  across long exons.

Although we cannot formally rule out this “relaxed constraints” possibility, several lines of evidence would argue against it. First, the relaxed constraints scenario would be associated with the length of the CDS, regardless of the exon-intron structure of the gene. As indicated above, intragenic studies of codon bias in *Drosophila* reveal differences between genes with and without introns and a U-shaped distribution of codon bias only observed in genes without introns (Kliman and Eyre-Walker 1998; Comeron and Kreitman 2002). Second, the relaxed constraints scenario predicts that this trend will be observed in all species while IS acting on intragenic patterns is expected to be sensitive to differences in recombination rates. A recent study of intragenic spatial patterns of codon bias in prokaryotes and eukaryotes reveals that the U-shaped distribution of codon bias across transcripts is not observed in prokaryotes (Qin et al. 2004). The lack of evidence for IS at the intragenic level in prokaryotes is not surprising because organisms with a very reduced effective recombination rate will have all sites in a gene linked, hence there is no possibility for intragenic differences due to IS (Qin et al. 2004). Third, relaxed constraints at the protein level in the central region of long

transcripts would predict a strong increase in  $K_a$  coupled with increased  $K_s$  and reduced  $\gamma_s$ . Our results do not support this scenario, with similar  $K_a$  in genes with long and short CDS and no significant increase in  $K_a$  in the central regions of long exons (see above). Altogether, the data is best explained by the IS model, at least in *Drosophila*. Nevertheless, studies of polymorphism and divergence comparing genes with different exon-intron structures and substantially different rates of crossing-over will be important in appraising more precisely IS, its effects, and its limits.

#### Possible Influence of Biased Mismatch Repair

A proposed explanation for some variation in codon usage among genes is the mechanism of biased mismatch repair after gene conversion events (BGCR). Indeed, this neutral mechanism can mimic the evolutionary consequences of weak selection in association with the frequency of gene conversion events (Nagyaki 1983). In particular, in species where most favored codons end in G or C, like *D. melanogaster* (Akashi 1995; Duret and Mouchiroud 1999), biased mismatch repair towards G or C would generate the same effects as translational selection (Eyre-Walker 1993; Birdsall 2002; Marais 2003). For that reason, this mechanism has been proposed as a possible neutral explanation for a positive relationship between rates of crossing-over (and likely gene conversion) and codon bias in *D. melanogaster* (Kliman and Hey 1993; Comeron, Kreitman, and Aguade 1999; Hey and Kliman 2002). Under this scenario, estimates of  $\gamma_s$  would denote the end result of selection intensity and BGCR.

Although there is controversy on whether there is evidence of BGCR towards G or C in *Drosophila* (Kliman and Hey 2003; Marais, Mouchiroud and Duret 2003), BGCR could only influence our conclusions if genes with variable CDS length were not distributed randomly across the genome, somewhat associated with recombination rates. Our two sets of genes are located in genomic regions with equivalent rates of crossing-over in *D. melanogaster* ( $P > 0.65$ ). Moreover, a complete genome analysis in *D. melanogaster* also showed no significant association between the length of the CDS and measures of recombination (Hey and Kliman 2002). Therefore, BGCR is an unlikely cause for the observed difference between genes with short and long CDS. On the other hand, there is no reason to assume that BGCR could generate a symmetrical intragenic pattern, with less biased or reduced repair in the central region of long exons.

#### Possible Influence of Dominance and Nonequilibrium

A caveat to our previous estimates of  $\gamma_s$  is that these estimates are based on theoretical predictions that assume genic selection and MSD equilibrium. Estimates of selection based on a ratio polymorphism to divergence (e.g., McDonald-Kreitman test or  $rpd$ ) are fairly robust to deviations from ideal panmictic populations and nearly unbiased assuming genic selection, even when dominance or recessivity is complete (Weinreich and Rand 2000; Wakeley 2003; Williamson, Fledel-Alon, and Bustamante 2004). On the other hand, changes in either mutational tendencies

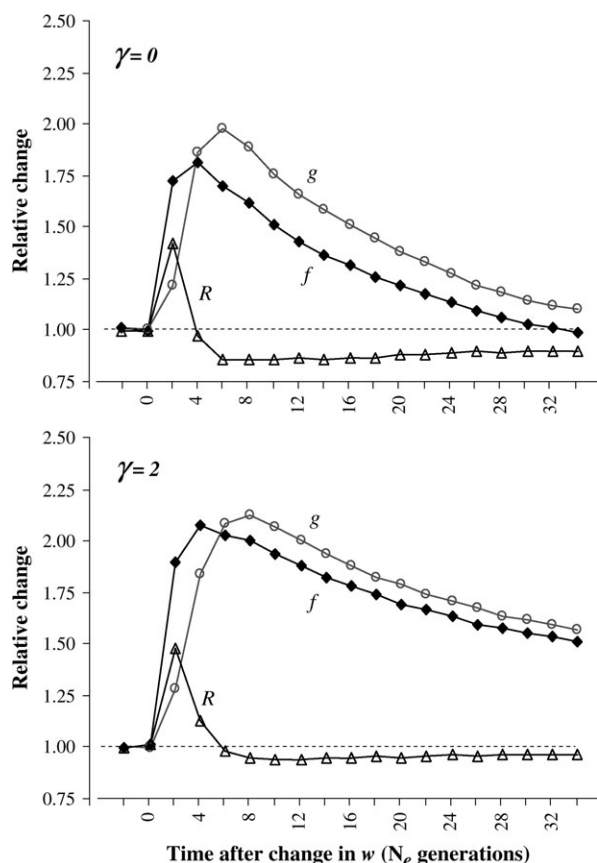


FIG. 3.—Influence of changing the mutation bias towards AT ( $w$ ). Relative influence of changing  $w$  from 0.61 to 0.77 on  $f$  (the ratio of U to P polymorphic mutations),  $g$  (the ratio of U to P fixed mutations), and  $R$  ( $ff/g$ ). Values for  $f$ ,  $g$  and  $R$  are relative to those at mutation-selection-drift equilibrium with  $w = 0.61$  and  $\gamma = 0$  or 2. Results based on 5,000 independent forward simulations comparing 500 sites evolving independently in two species with diploid population size of 10,000. Sample size ( $n$ ) is 10 chromosomes in each species. For simplicity, we assumed a biallelic model with P and U mutations being equivalent to GC and AT mutations.

or  $N_e$  could influence our estimates of  $\gamma_s$ . It is important to recall, however, that our results are based on a consistent difference between genes with short and long CDS and across long exons. It is most unlikely that departures from MSD equilibrium could generate a systematically different bias in the central region of long exons compared to the flanking regions, or in genes with long exons compared to genes with short exons, unless the initial conditions were already different.

We investigated the effects of a change in the mutation bias towards AT ( $w$ ) on estimates of  $\gamma_s$  with forward simulations. Following Kern and Begun (2005) we studied a change in  $w$  from 0.61 to 0.77 proposed for *D. melanogaster* lineage. Figure 3 shows the influence of increasing  $w$  on  $f$ ,  $g$  (the ratio of U to P fixed mutations), and  $R$  ( $ff/g$ , a measure positively correlated with  $\gamma_{s-rpd}$ ). As expected,  $f$  is strongly influenced by changes in  $w$ . Any increase in  $w$  will cause a fast increase in  $f$  and an overestimation of  $\gamma_{s-f}$  that is persists for many  $N_e$  generations. On the other hand,  $R$  first increases, causing an overestimate of  $\gamma_{s-rpd}$ , to decrease after  $\sim 4N_e$  generations, reaching values smaller than those at

equilibrium, hence underestimating  $\gamma_{s-rpd}$ . In all, changes in either  $w$  or  $N_e$  will influence our estimates of  $\gamma_s$ , mainly with opposite biases for  $\gamma_{s-rpd}$  and  $\gamma_{s-f}$ . Therefore, a change in  $w$  is an unlikely explanation to a pattern observed with both measures of  $\gamma_s$ .

#### Influence of *D. melanogaster* Lineage

Several lines of evidence suggest that *D. melanogaster* lineage may not be at MSD equilibrium. Analyses of P and U mutations in the *D. melanogaster* and *D. simulans* lineages suggest a relaxation of selection or an increased  $w$  in the *D. melanogaster* lineage (Akashi 1995; Akashi 1996; Akashi 1997; Eyre-Walker 1997; Takano-Shimizu 2001). More recent analyses of intergenic and intron sequences also support an increase in  $w$  in *D. melanogaster*, while no mutational change is detected in the *D. simulans* lineage (Kern and Begun 2005). We then estimated  $\gamma_{s-rpd}$  and  $\gamma_{s-f}$  removing *D. melanogaster* lineage data (see Table 3). Estimates of  $\gamma_{s-rpd}$  increase for all genes, with  $\gamma_{s-rpd}(\text{short CDS}) = 2.06$  and  $\gamma_{s-rpd}(\text{long CDS}) = 1.58$ . We also detect that  $\gamma_{s-f}$  is somewhat reduced when the *D. melanogaster* data is removed from the study, with  $\gamma_{s-f}(\text{short CDS}) = 2.94$  and  $\gamma_{s-f}(\text{long CDS}) = 1.68$ . The detection of an increase in  $\gamma_{s-rpd}$  and a decrease in  $\gamma_{s-f}$  when *D. melanogaster* lineage data is removed is in agreement with the expected biases observed in our simulations. In all,  $\gamma_s$  in long exons remains significantly smaller than that observed in short exons under both approaches ( $P < 0.0001$  and  $P = 0.0032$  for  $\gamma_{s-rpd}$  and  $\gamma_{s-f}$ , respectively). Estimates of  $\gamma_s$  across long exons removing *D. melanogaster* data also show reduced  $\gamma_s$  in the central region.  $\gamma_{s-rpd}$  is 2.58, 1.46, and 1.84 for the three nonoverlapping 5', central, and 3' regions, respectively;  $\gamma_{s-rpd}$  in the central region is significantly smaller than that in the 5' ( $P = 0.0003$ ) or 3' ( $P = 0.029$ ) flanking regions. Estimates of  $\gamma_{s-f}$  are 2.55, 1.53, and 2.34 for the 5', central, and 3' regions, respectively.  $\gamma_{s-f}$  in the central region is significantly smaller than  $\gamma_{s-f}$  in the flanking regions ( $P = 0.0005$  and  $P = 0.0021$  when comparing to 5'- and 3'-flanking regions, respectively) or when compared to both flanking regions combined ( $P < 1 \times 10^{-5}$ ). Thus, our analyses indicate that the inclusion of *D. melanogaster* data influences our precise estimates of  $\gamma_s$ , but it is neither responsible for the observation that  $\gamma_s$  is significantly greater than zero nor for the observed difference in  $\gamma_s$  between genes with short and long CDS or across long exons.

#### Conclusions

The IS model proposes that the HR effect has detectable evolutionary consequences at the intragenic level in species with recombination, with exon size, and exon-intron structure playing a role in shaping the effectiveness of selection. Here, we examined two specific predictions of this model: (1) long exons will show reduced effectiveness of selection when compared to small exons, and (2) the central region of long exons will show reduced intensity of selection compared to flanking coding regions of the same exons. We investigated rates of synonymous evolution ( $K_s$ ) and selection intensity on synonymous mutations ( $\gamma_s$ ) in genes with similar levels of expression and rates of crossing-over in

*D. melanogaster*. Our results reveal that synonymous mutations in genes with long exons show higher  $K_s$  and reduced  $\gamma_s$  than those in genes with short exons.  $K_s$  and  $\gamma_s$  also vary significantly across long exons, with increased  $K_s$  and reduced  $\gamma_s$  in the central region compared to flanking regions of the same exons. This latter observation suggests that the difference between genes with short and long exons can be mostly attributed to the central region of these long exons. All these results are consistent with predictions of the IS model. The use of polymorphism and divergence data allows us to rule out possible differences in mutational rates or patterns among or across genes. The comparison of rates of nonsynonymous evolution ( $K_a$ ) indicates that the observed differences among and across exons cannot be explained by differences in protein constraints. Finally, the study of  $K_s$  and  $\gamma_s$  separately for each amino acid also allows us to rule out any possible influence of protein composition.

Our results further stress that most synonymous mutations will not provide a proper measure of neutral mutation rates in species such as *Drosophila*, with detectable selection on codon usage (see McVean and Vieira 2001, Dumont et al. 2004). Studies on protein evolution based on comparing different genes, different regions of the same CDS, or even different amino acids that use synonymous mutations should be interpreted bearing in mind not only the possibility of selection on synonymous mutations but also, and more importantly, the consequences of variation in  $\gamma_s$ .

### Acknowledgments

We thank Ana Llopart for valuable comments and discussion. We also thank two anonymous reviewers and Marta Wayne for their helpful comments and Carlos D. Bustamante and CBSU for the program mkprf. This work was supported by U.S. National Science Foundation grant DEB-03-44209 to J.M.C.

### Literature Cited

- Adams, M., S. Celniker, R. Holt et al. (95 co-authors). 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.
- . 1995. Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* **139**:1067–1076.
- . 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.
- . 1997. Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. *Genetics* **147**:1989–1991.
- . 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**:221–238.
- . 2003. Translational selection and yeast proteome evolution. *Genetics* **164**:1291–1303.
- Akashi, H., and S. W. Schaeffer. 1997. Natural selection and the frequency distributions of “silent” DNA polymorphism in *Drosophila*. *Genetics* **146**:295–307.
- Ashburner, M. 1989. *Drosophila: a laboratory handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- Barrier, M., C. D. Bustamante, J. Yu, and M. D. Purugganan. 2003. Selection on rapidly evolving proteins in the Arabidopsis Genome. *Genetics* **163**:723–733.
- Barton, N. 1995. Linkage and the limits to natural selection. *Genetics* **140**:821–841.
- Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **99**:13616–13620.
- Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19**:1181–1197.
- Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**:897–907.
- Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2003. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* **63**:91–103.
- Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in Arabidopsis. *Nature* **416**:531–534.
- Caballero, A., and E. Santiago. 1995. Response to selection from new mutation and effective size of partially inbred populations. I. Theoretical results. *Genet. Res.* **66**:213–225.
- Carlini, D. B., Y. Chen, and W. Stephan. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics* **159**:623–633.
- Carlini, D. B., and W. Stephan. 2003. In vivo introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. *Genetics* **163**:239–243.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**:1289–1303.
- Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA* **101**:3480–3485.
- Coghlan, A., and K. H. Wolfe. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**:1131–1145.
- Comeron, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* **41**:1152–1159.
- . 1999. K-estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15**:763–764.
- . 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**:1293–1304.
- Comeron, J. M., and M. Kreitman. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**:389–410.
- Comeron, J. M., M. Kreitman, and M. Aguade. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* **151**:239–249.
- Crow, J. F., and K. Kimura. 1970. *An introduction to population genetics theory*. Harper and Row, New York.
- Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**:205–216.

- DuMont, V. B., J. C. Fay, P. P. Calabrese, and C. F. Aquadro. 2004. DNA variability and divergence at the Notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167**:171–185.
- Duret, L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16**:287–289.
- 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.
- Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. Lond. B* **252**:237–243.
- . 1997. Differentiating between selection and mutation bias. *Genetics* **147**:1983–1987.
- Fay, J. C., G. J. Wyckoff, and C. I. Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**:1024–1026.
- Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics* **78**:737–756.
- FLYBASE. 1998. FlyBase—a *Drosophila* database. *Nucleic Acids Res.* **26**:85–88.
- Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- Hey, J., and R. M. Kliman. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* **160**:595–608.
- 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.
- Kern, A. D., and D. J. Begun. 2005. Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. *Mol. Biol. Evol.* **22**:51–62.
- Kimura, M. 1964. Diffusion models in population genetics. *J. Appl. Prob.* **1**:177–232.
- . 1981. Estimation of evolutionary distances between homologous nucleotide Sequences. *Proc. Natl. Acad. Sci. USA* **78**:454–458.
- Kliman, R. M., and A. Eyre-Walker. 1998. Patterns of base composition within the genes of *Drosophila melanogaster*. *J. Mol. Evol.* **46**:534–541.
- Kliman, R. M., and J. Hey. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**:1239–1258.
- . 2003. Hill–Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. *Genet. Res.* **81**:89–90.
- Kliman, R. M., N. Irving, and M. Santiago. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* **57**:98–109.
- Kondrashov, A. S. 1994. Muller’s ratchet under epistatic selection. *Genetics* **136**:1469–1473.
- Kreitman, M., and M. A. Antezana. 2000. The population and evolutionary genetics of codon bias. Pp. 82–101 in R. S. Singh, and C. B. Krimbas, eds. *Evolutionary genetics: from molecules to morphology*. Cambridge University Press, Cambridge.
- Li, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**:337–345.
- . 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**:96–99.
- Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19**:330–338.
- Marais, G., D. Mouchiroud, and L. Duret. 2003. Neutral effect of recombination on base composition in *Drosophila*. *Genet. Res.* **81**:79–87.
- Maside, X., A. W. Lee, and B. Charlesworth. 2004. Selection on codon usage in *Drosophila americana*. *Curr. Biol.* **14**:150–154.
- 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. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74**:145–158.
- . 2000. The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**:929–944.
- McVean, G. A., and J. Vieira. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**:245–257.
- 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. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* **26**:3188–3193.
- Nagylaki, T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* **80**:6278–6281.
- Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929–936.
- Otto, S. P., and N. H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. *Genetics* **147**:879–906.
- Piganeau, G., R. Westrelin, B. Tourancheau, and C. Gautier. 2001. Multiplicative versus additive selection in relation to genome evolution: a simulation study. *Genet. Res.* **78**:171–175.
- Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. *Proc. Natl. Acad. Sci. USA* **101**:12588–12591.
- Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **94**:7784–7790.
- Powell, J. R., E. Sezzi, E. N. Moriyama, J. M. Gleason, and A. Caccone. 2003. Analysis of a shift in codon usage in *Drosophila*. *J. Mol. Evol.* **57**(Suppl. 1):S214–S225.
- Qin, H., W. B. Wu, J. M. Comeron, M. Kreitman, and W. H. Li. 2004. Intra-genic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* **168**:2245–2260.
- Sawyer, S. A., and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. *Genetics* **132**:1161–1176.
- Sharp, P. M., and W. H. Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**:28–38.
- . 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
- Smith, N. G., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. *Nature* **415**:1022–1024.
- Stolc, V., Z. Gauhar, C. Mason et al. (12 co-authors). 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**:655–660.
- Tachida, H. 2000. Molecular evolution in a multisite nearly neutral mutation model. *J. Mol. Evol.* **50**:69–81.
- Takano-Shimizu, T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. *Mol. Biol. Evol.* **18**:606–619.
- Wakeley, J. 2003. Polymorphism and divergence for island-model species. *Genetics* **163**:411–420.
- Weinreich, D. M., and D. M. Rand. 2000. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics* **156**:385–399.

Williamson, S., A. Fledel-Alon, and C. D. Bustamante. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **168**:463–475.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.

Marta Wayne, Associate Editor

Accepted August 15, 2005