Molecular Biology and Evolution

Patterns of Mutation and Selection at Synonymous Sites in *Drosophila*

Nadia D. Singh, Vanessa L. Bauer DuMont,* Melissa J. Hubisz,† Rasmus Nielsen,‡ and Charles F. Aquadro**

*Department of Molecular Biology and Genetics, Cornell University, Ithaca; †Department of Human Genetics, University of Chicago; and ‡Institute of Biology and Centre for Bioinformatics, University of Copenhagen, Copenhagen, Denmark

That natural selection affects molecular evolution at synonymous sites in protein-coding sequences is well established and is thought to predominantly reflect selection for translational efficiency/accuracy mediated through codon bias. However, a recently developed maximum likelihood framework, when applied to 18 coding sequences in 3 species of *Drosophila*, confirmed an earlier report that the *Notch* gene in *Drosophila melanogaster* was evolving under selection in favor of those codons defined as unpreferred in this species. This finding opened the possibility that synonymous sites may be subject to a variety of selective pressures beyond weak selection for increased frequencies of the codons currently defined as "preferred" in *D. melanogaster*. To further explore patterns of synonymous site evolution in *Drosophila* in a lineage-specific manner, we expanded the application of the maximum likelihood framework to 8,452 protein coding sequences with well-defined orthology in *D. melanogaster*, *Drosophila sechellia*, and *Drosophila yakuba*. Our analyses reveal intragenomic and interspecific variation in mutational patterns as well as in patterns and intensity of selection on synonymous sites. In *D. melanogaster*, our results provide little statistical evidence for recent selection on synonymous sites, and *Notch* remains an outlier. In contrast, in *D. sechellia* our findings provide evidence in support of selection predominantly in favor of preferred codons. However, there is a small subset of genes in this species that appear to be evolving under selection in favor of unpreferred codons, which indicates that selection on synonymous sites is not limited to the preferential fixation of mutations that enhance the speed or accuracy of translation in this species.

Introduction

Molecular evolution at synonymous sites is thought to be largely characterized by selection for biased codon usage in *Drosophila*, yeast, and bacteria. The frequent presence of major codons, which are thought to correspond to the most abundant tRNAs, in protein coding sequences is thought to improve the accuracy and/or the efficiency of translation (Bulmer 1991; Akashi and Eyre-Walker 1998; Akashi et al. 1998). In *Drosophila*, there does appear to be a well-defined set of "major" or "preferred" codons. These definitions are based on identifying those codons that increase in frequency between genes with low codon bias versus high codon bias using correspondence analysis or other similar techniques. Although there are subtle differences in the precise definition of the preferred codons, it is generally accepted that with the exception of at most one codon, preferred codons in *Drosophila melanogaster* are G or C ending (Akashi 1995; Duret and Mouchiroud 1999). Moreover, these codon preferences appear to be generally conserved across many *Drosophila* species; major codons between *D. melanogaster* and *Drosophila pseudoobscura*, for instance, are nearly identical (Akashi 1997).

The evolution of codon bias and the nature of selection on biased codon usage in *Drosophila* have been extensively investigated. In general, the estimated strength of selection on codon bias in *Drosophila* is quite weak (Akashi 1995, 1997; McVean and Vieira 2001; Singh et al. 2004), but because codon bias is believed to be maintained by mutation–selection–drift equilibrium (Sharp and Li 1986; Bulmer 1991; Akashi and Schaeffer 1997; McVean and Charlesworth 1999), there is substantial variation in the degree of codon bias within and between genomes. For instance, within a given genome, codon bias is known to be correlated with genic features such as protein length (Akashi 1996; Eyre-Walker 1996; Comeron et al. 1999; Duret and Mouchiroud 1999; Marais and Duret 2001), recombination rate (Kliman and Hey 1993; Comeron et al. 1999; Marais et al. 2001, 2003; Hey and Kliman 2002), rate of protein evolution (Akashi 1996; Cutter et al. 2003), expression level (Sharp and Li 1986; see also Bulmer 1988; Duret and Mouchiroud 1999; reviewed in Akashi 2001; Hey and Kliman 2002), gene density (Hey and Kliman 2002), chromosomal location (Comeron et al. 1999; Hambuch and Parsch 2005; Singh et al. 2005b), as well as features such as the presence and length of introns (Comeron and Kreitman 2000; Vinogradov 2001).

In addition to varying intragenomically, codon bias also varies between species. The *D. melanogaster* lineage, for instance, appears to have undergone a dramatic reduction in codon bias since its divergence with *Drosophila simulans* (Akashi 1995, 1996; McVean and Charlesworth 1999; Bauer DuMont et al. 2004; Akashi et al. 2006; Nielsen et al. 2007). In addition, patterns of synonymous codon usage in the *Drosophila saltans/Drosophila willistoni* clade appear to deviate significantly from that of its relatives in the *melanogaster* and *obscura* groups in the subgenus *Drosophila* (Anderson et al. 1993; Rodriguez-Trelles et al. 1999a, 1999b, 2000a, 2000b; Powell et al. 2003).

Because codon bias is maintained by mutation–selection–drift balance, these intragenomic and interspecific heterogeneities in patterns of codon usage may result from differences in the strength of selection on codon bias, differences in effective population size (mediated at least in part by both intragenomic and interspecific variation in recombination rate), as well as differences in mutation biases. Disentangling the roles of mutation versus the strength and efficacy of selection in the evolution of codon bias has proven challenging thus far. Recently, a maximum likelihood framework was developed which estimates not only patterns of single nucleotide mutations but also the strength of selection on synonymous sites (Nielsen et al. 2007). Importantly, this methodology allows for lineage-specific parameter estimation, which facilitates interspecific comparisons of both mutation rates and the nature of selection on synonymous sites.

Key words: synonymous site, codon bias, mutational patterns.

E-mail: nds25@cornell.edu.

Mol. Biol. Evol. 24(12):2687–2697. 2007
doi:10.1093/molbev/msm196
Advance Access publication September 25, 2007

http://mbe.oxfordjournals.org/

The results from the initial application of this method to 18 protein-coding sequences (Nielsen et al. 2007) hinted at the possibility of interspecific differences in mutation rates between *D. melanogaster* and its sister species *D. simulans* and were suggestive of differences in patterns of selection at synonymous sites between species. Whereas in *D. simulans*, these results provided support for selection in favor of preferred codons in several genes, only 1 of the 18 genes in *D. melanogaster*, *Notch*, showed signs of selection at synonymous sites. Consistent with previous results (Bauer DuMont et al. 2004), *Notch* was shown to be evolving under selection in favor of unpreferred codons in this lineage, which clearly deviates from the traditional model of selection on synonymous sites wherein preferred codons are selectively favored.

The results from this initial study (Nielsen et al. 2007) yielded 3 unresolved questions. One, is there any evidence for significant differences in mutation rates among closely related *Drosophila* species at a genomic scale? Two, in addition to *Notch*, are there other genes in *Drosophila* that show evidence of selection toward unpreferred codons? Three, do patterns of synonymous site evolution differ interspecifically across the genome? Here, we present an application of this maximum likelihood framework to a set of 8,452 genes aligned among *D. melanogaster*, *Drosophila sechellia*, and *Drosophila yakuba*. This application facilitated estimating lineage-specific mutation rates as well as lineage-specific gene-by-gene estimates of the selection parameter on synonymous sites ($S = 2N_e s$) for both *D. melanogaster* and *D. sechellia*. Note that for the current application of this methodology, we chose to use coding sequences from *D. sechellia* instead of *D. simulans* given the apparently slightly higher quality of the data in the former (see Methods); the results from *D. sechellia* are likely to be comparable to those from *D. simulans* given that the majority of their evolutionary history has been shared. Moreover, distributions of codon bias show no significant differences between *D. sechellia* and *D. simulans* (data not shown), which suggests that any trend specific to one of these lineages is likely to have evolved quite recently and would thus not have marked impact on our results.

Our results are suggestive of differences in mutational patterns between the X and the autosomes within each species and also indicate clear differences between species with respect to the relative rates of each of the 12 single nucleotide mutations. Moreover, our results suggest that in *D. sechellia* and *D. melanogaster*, whereas the majority of genes under selection at synonymous sites show selection in favor of preferred codons, there does exist a small subset of genes evolving under selection toward unpreferred codons. These results represent the first documentation of selection in favor of unpreferred codons in an additional Drosophilid, which suggests that patterns of synonymous site evolution at *Notch* in *D. melanogaster* are not unique. However, very few genes in *D. melanogaster* show evidence for selection after correction for multiple tests, whereas this is not the case for *D. sechellia*. Thus, our analyses highlight the dynamic nature of patterns of mutation and selection at synonymous sites on an evolutionary timescale.

Methods

Coding Sequences

We downloaded all of the coding sequences that have been aligned in the *D. melanogaster* species group from the AAA Web site (http://rana.lbl.gov/~venky/AAA/freeze_20061030/protein_coding_gene/GLEANR/alignment/), which are presented elsewhere (Drosophila 12 Genome Consortium, forthcoming). We used the masked alignments of single-copy orthologs based on the longest coding sequences and with a guide tree. From these 6 species alignments, we extracted those sequences corresponding to those from *D. melanogaster*, *D. sechellia*, and *D. yakuba* and stripped the alignments of sites that contained bases that were masked (details on the masking protocol can be found at http://rana.lbl.gov/drosophila/wiki/index.php/Alignment_Masking) in at least one of these species (Sackton T, Larracuente A, personal communication). We chose to use *D. sechellia* instead of *D. simulans* because *D. sechellia* had a lower proportion of masked sites within the multispecies coding sequence alignments than *D. simulans*. This procedure resulted in 8,452 coding sequences that were aligned between *D. melanogaster*, *D. sechellia*, and *D. yakuba*.

Given the potential for sex-specific mutational biases, we separated X-linked and autosomal genes for our analysis. Chromosomal locations were taken from the gene annotations in Release 4.3 of the *D. melanogaster* genome, and we assumed that the chromosomal locations of the orthologous sequences in *D. sechellia* and *D. yakuba* were the same. Although there are likely to be some genes that have moved from one Muller element to another in these lineages, we expect that this number will be very small given the low rate of interchromosomal gene movement in *Drosophila* (Ranz et al. 2001; Richards et al. 2005) and thus would not significantly affect our results.

Maximum Likelihood Estimation of Selection on Codon Usage

To estimate the selection for optimal codon usage bias on each lineage of the phylogeny, we use the method of Nielsen et al. (2007). In brief, this method is an extension of the classical codon-based likelihood models (Goldman and Yang 1994; Muse and Gaut 1994) that allows selection for optimal codon usage. S, the scaled selection coefficient in favor of mutations from the unpreferred to the preferred state, is a parameter of the Markov model and is estimated independently on each lineage of the tree. If S is positive, this implies selection in favor of the preferred state, and if S is negative, this implies selection against the preferred state. A likelihood ratio test of the null hypothesis of no selection, H_0: $S = 0$, is performed by comparing the maximum likelihood values under a model in which $S = 0$ to a model in which S is considered a free parameter. In addition to S, parameters of the mutation matrix and ω, the d_N/d_S ratio, are also estimated on each lineage.

Other Genic Features

Two metrics of codon bias were calculated for each of the 8,452 coding sequences in our analysis. Both the

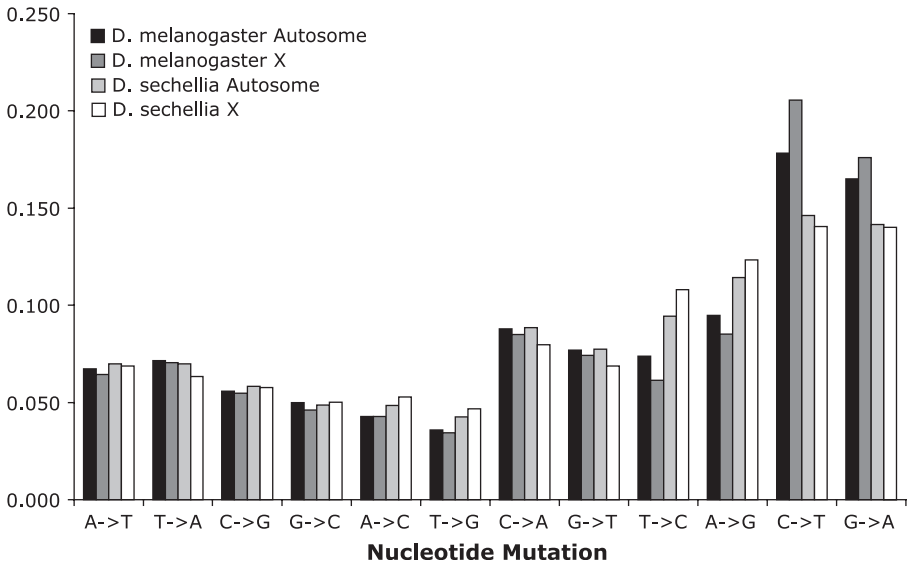

FIG. 1.—Relative rates of each of the 12 single nucleotide mutations in autosomal and X-linked genes of *Drosophila melanogaster* and *Drosophila sechellia*.

frequency of optimal codons (FOP) (Ikemura 1981) and the effective number of codons (ENC) (Wright 1990) were estimated using codonW (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html). For the estimation of FOP, the optimal codons for *D. melanogaster* were used for all species, which is appropriate given that codon preferences appear to be quite conserved across *Drosophila* (Akashi and Schaeffer 1997).

We also retrieved information on the length of the coding sequences for *D. sechellia* for all of the genes included in our analyses. These data were extracted from the gff files for each species (http://rana.lbl.gov/~venky/AAA/freeze_20061030/protein_coding_gene/GLEANR/annotation/) (Larracuente A, personal communication).

GC Content

Equilibrium GC content (π_{GC}) was calculated as the stationary distribution of the mutation rate matrix. The equilibrium frequency is found for each base by solving a system of 4 equations such as $(q_{A\to C} + q_{A\to T} + q_{A\to G})\ \pi_A = q_{T\to A}\ \pi_T + q_{G\to A}\ \pi_G + q_{C\to A}\ \pi_C$, where $q_{i\to j}$ is the mutation rate from base i to base j. We then calculate $\pi_{CG} = \pi_G + \pi_C$. Observed GC content was calculated from 4-fold degenerate sites along the entire aligned gene sequence in *D. melanogaster* and *D. sechellia*.

Gene Ontology

We investigated whether certain functional categories of genes had characteristically elevated estimated strengths of selection. Given the samples sizes of gene subsets, we only considered genes with positive selection parameter estimates in *D. sechellia*. We tested whether genes associated with given gene ontology (GO) terms had an elevated median strength of selection relative to expectation given the distribution of the selection parameters for the gene study set. The GO terms used were based on a customized *D. melanogaster* GO-SLIM association file (Larracuente A, Sacton T, personal communciation). To assess significance, we used permutation tests; data sets were permuted 10,000 times to calculate the *P* value, following procedures described previously (Drosophila 12 Genomes Consortium, forthcoming).

We also used the Web implementation of GeneMerge (http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge/GeneMerge.html) (Castillo-Davis and Hartl 2003) to test whether there were GO terms that were significantly over- or underrepresented in our 5% false-discovery rate (FDR) gene set.

Results

Mutation Parameter Estimation

We estimated the relative rates of each of the 12 single nucleotide mutations in *D. melanogaster* and *D. sechellia* for X-linked and autosomal genes; these data are presented in figure 1. For this global likelihood analysis, all of the coding sequences were concatenated together for each species to generate lineage-specific mutation rate estimates (see Methods). Because of the difficulty in precisely placing the root on this 3-species phylogeny, we cannot accurately estimate the rates of these mutations in the *D. yakuba* lineage. In both *D. melanogaster* and *D. sechellia*, the rates of C→T and G→A transitions are consistently higher than the rates of all other mutations for both X-linked and autosomal genes. Interestingly, the rates of A→G and T→C mutations appear to be increased in *D. sechellia* in both X-linked and autosomal genes relative to *D. melanogaster*. Additionally, the relative rates of the 4 transitions seem to differ on the X and the autosomes, though not consistently between species. Although in *D. melanogaster* the C→T and G→A mutations are increased on the X relative to

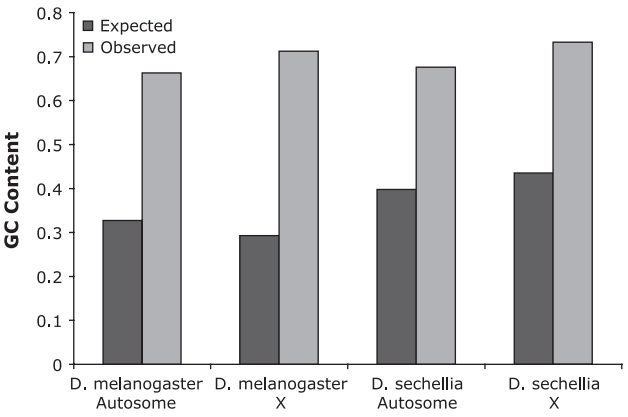

FIG. 2.—Expected versus observed (GC4) content of X-linked and autosomal genes in *Drosophila melanogaster* and *Drosophila sechellia*.

the autosomes, the relative rates of these mutations appear to be similar on the X and the autosomes in *D. sechellia*. In addition, whereas the rates of the A→G and T→C mutations are higher on the autosomes than on the X in *D. melanogaster*, the opposite is true in *D. sechellia* (fig. 1).

To assess the overall impact of these variations in the relative rates of each of the 12 single nucleotide mutations, we calculated equilibrium GC content (π_{GC}) as the stationary distribution of the mutation matrix (see Methods). These data are presented as the "expected" GC content in figure 2. In both species, there do appear to be differences in the mutational profiles of X-linked and autosomal genes. In *D. melanogaster*, π_{GC} is estimated as 32.7% and 29.3% for autosomal and X-linked genes, respectively, whereas in *D. sechellia*, π_{GC} is estimated at 39.8% and 43.4% for autosomal and X-linked genes, respectively. We cannot currently assess whether these intragenomic differences between X-linked and autosomal sequences are statistically significant given our methodology, but this will be pursued further in the future.

These intragenomic differences in π_{GC} between X-linked and autosomal genes are echoed in the interspecific comparisons of the mutational spectra between *D. melanogaster* and *D. sechellia* in a likelihood framework. The mutation profiles of X-linked genes and autosomal genes are significantly different between the 2 species ($P << 0.0001$, likelihood ratio test, both comparisons). For both autosomal and X-linked genes, the relative rates of the transitions appear to be those that differ most dramatically between the 2 species. In particular, the rates of A→G and T→C appear to be higher in *D. sechellia* than in *D. melanogaster*, and the rates of the reverse mutations, the C→T and G→A transitions, appear to be lower in *D. sechellia* (fig. 1), which leads to an increase in the estimate of π_{GC} in *D. sechellia* versus *D. melanogaster* (fig. 2).

We also compared the inferred equilibrium GC content with the observed GC content for X-linked and autosomal genes in both species. In all comparisons, the observed GC content is substantially increased relative to expectation and the effect appears to be exaggerated in *D. melanogaster* (fig. 2). In addition, whereas in *D. sechellia* both observed and expected GC content are increased on the X chromosome relative to the autosomes, in *D. melanogaster* the expected GC content is decreased on the X chromosome while the observed GC content is increased in relation to the autosomes.

Selection Parameter Estimation

Selection parameters ($S = 2N_e s$) were estimated in a gene-by-gene analysis using the species-specific mutation rates inferred by the global likelihood analysis presented above. Estimates of $S > 0$ are indicative of selection in favor of preferred codons, whereas estimates of $S < 0$ are consistent with selection in favor of unpreferred codons. We will refer to estimates of S in *D. melanogaster* as "S_m" and in *D. sechellia* as "S_s". Because these gene-by-gene analyses are performed within a likelihood framework, we can restrict our analyses to those genes with some evidence of selection on synonymous sites. Namely, we can focus on those genes that have at least 5 synonymous substitutions and also significantly reject the null hypothesis that there is no selection on synonymous sites ($S = 0$). There were 1,609 genes for which the null model was rejected in at least one species and 138 genes that significantly reject the null hypothesis in both species. In *D. melanogaster*, there were 408 and 171 genes with $S_m > 0$ and $S_m < 0$, respectively, and there were, respectively, 941 and 136 genes with $S_s > 0$ and $S_s < 0$ in *D. sechellia*.

It is important to note, however, that although the separation of X-linked and autosomal genes for estimating mutation matrices does capture some of the intragenomic variation in mutation rates, it remains possible that mutation rates vary even within these chromosome sets (Singh et al. 2004). The use of a single set of mutation rates for each chromosome set may thus influence our estimates of selection parameters. To investigate this possibility, we ran the maximum likelihood model on the pooled autosomal and X-linked genes, which resulted in an estimation of a genomic mutation rate matrix as well as gene-by-gene estimates of S. Reasoning that autosomal genes would principally contribute to the genomic mutation matrix given the relative abundance of autosomal and X-linked genes in the genome, we compared the selection parameter estimates of X-linked genes based on the genomic mutation matrix with those estimated using a X chromosome–specific mutation matrix.

Overall, estimates of S_m and S_s of X-linked genes were highly correlated between the 2 models (*D. melanogaster* Kendall's $\tau = 0.87$, $P = 0.0004$; *D. sechellia* Kendall's $\tau = 0.95$, $P << 0.0001$). However, in *D. melanogaster*, there were 67 genes with significant evidence for selection on synonymous sites based on the genomic mutation matrix that did not show evidence for selection when the X chromosome mutation matrix was employed, all of which had negative selection parameter estimates. In addition, there were 32 genes with evidence of selection based on the X chromosome mutation matrix that did not appear to be under selection when the genomic mutation matrix was employed; 31 of these genes had positive selection parameter estimates, whereas the remaining gene had a negative selection parameter estimate. The results from *D. sechellia* were much more consistent between the 2 models. There were 10 genes showing selection based on the genomic

Table 1
Genes in *Drosophila melanogaster* with Evidence for Selection on Synonymous Sites

Gene	Chromosome	S	q value
CG2493	2L	3.07	0.007
Notch	X	−1.09	0.015
lava lamp	X	1.08	0.037

mutation matrix that did not appear to be under selection when the X chromosome mutation matrix was used, 9 of which had positive selection parameter estimates. In addition, of the 3 genes with significant evidence of selection with the X chromosome mutation matrix that did not show evidence of selection with the genomic mutation matrix, all 3 had negative selection parameter estimates.

It is thus clear that although the quantitative estimates of selection seem comparatively robust to variation in the mutation rate matrices, the enumeration of genes with statistical support for selection on synonymous sites appears slightly more sensitive, particularly in *D. melanogaster*. The precise effect of masking heterogeneity in mutation rates on selection parameter estimation depends on the differences in GC bias between the more general and more precise mutation matrices. For instance, in *D. melanogaster*, the X chromosome mutation matrix is more AT biased than the genomic mutation matrix (data not shown), and as a consequence, the application of the genomic mutation matrix yielded several genes with perhaps erroneous support of selection favoring the fixation of AT-rich unpreferred codons. In contrast, in *D. sechellia*, the X chromosome matrix is more GC biased than the genomic mutation matrix, and as a result, employing the genome mutation matrix yielded a few likely erroneous genes with evidence of selection in favor of GC-rich preferred codons. Further, more precise specification of the mutation rate matrix in *D. melanogaster* provided additional genes with evidence of selection with predominantly positive selection parameter estimates, whereas in *D. sechellia*, this led to additional genes with statistically significant negative selection parameter estimates. Future analyses where mutation matrices derived from adjacent noncoding sequences are examined on gene-by-gene basis will help refine our inferences of precisely which genes show strong evidence for selection on synonymous sites and the estimates of the selection parameter.

Given the effects of oversimplification of the mutation matrix on our inferences of selection, coupled with the fact that we have performed over 8,000 tests, there is little question that some of the genes with evidence for selection at synonymous sites correspond to false positives. A 5% false discovery rate (FDR) set (Storey 2002) contains 3 genes in *D. melanogaster* (1 with $S_m < 0$ and 2 with $S_m > 0$, table 1) and 238 genes in *D. sechellia* (8 with $S_s < 0$, table 2, and 230 with $S_s > 0$, table 3). Because we are interested in identifying those genomic features that appear to be associated with selection on synonymous sites in these species, the more restricted data set with a stringent false-discovery threshold is most appropriate. As a consequence, the bulk of the analyses presented are based on those genes with positive estimates of S in *D. sechellia* as there are too few genes in *D. melanogaster* with evidence of selection and too few genes with negative estimates of S in *D. sechellia* for comprehensive analysis. Where appropriate, we also draw on comparisons between genes in the 5% FDR set and the full gene set to elucidate characteristics of those genes that show evidence of selection on synonymous sites.

Table 2
Genes in *Drosophila sechellia* with Evidence for Selection on Synonymous Sites toward Unpreferred Codons

Gene	Chromosome	S	q value
kurtz	3R	−8.33	0.005
CG32121	3L	−2.97	0.010
CG17471	4	−4.54	0.025
ebi	2L	−6.85	0.026
CG5149	2L	−5.11	0.036
Inhibitor-2	3L	−29.98	0.043
Rrp46	3R	−5.63	0.046
Suv4-20	X	−1.25	0.050

Genes with estimates of $S > 0$ may be under selection for increased codon bias wherein preferred codons are selectively favored. We therefore examined the relationships between several genic parameters relating to codon bias and the selection parameter estimate for this subset of genes in *D. sechellia*. X-linked and autosomal loci have been combined because there did not appear to be consistent differences between the chromosome sets either with respect to the distribution of the selection parameter estimates for X-linked versus autosomal genes or in relation to the relationships with other features. Consistent with expectation, genes with $S > 0$ show significant increases in the selection parameter with increased codon usage bias, as S is significantly correlated with both FOP (Kendall's $\tau = 0.24$, $P < 0.0001$) and ENC (Kendall's $\tau = -0.22$, $P < 0.0001$). Given that codon bias and protein length are negatively

Table 3
Top 20 Genes in *Drosophila sechellia* with Evidence for Selection on Synonymous Sites toward Preferred Codons

Gene	Chromosome	S	q value
Notch	X	2.51	<0.001
CG14651	3R	3.18	<0.001
scarface	2R	3.07	<0.001
Odorant receptor 42b	2R	4.29	<0.001
Tiggrin	2L	1.54	<0.001
CG9319	2L	2.89	<0.001
Polypeptide N-acetylgalactosaminyl transferase 35A	2L	3.09	<0.001
CG4825	3L	3.44	<0.001
Cyp303a1	2L	30.00	<0.001
CG14639	3R	3.18	<0.001
pollux	3R	1.58	<0.001
CG31720	2L	2.51	0.001
CG2254	X	3.07	0.001
CG1113	3R	2.20	0.001
CG2493	2L	5.47	0.001
CG14339	2L	2.06	0.001
CG7856	2R	2.86	0.002
Neprilysin 3	X	2.49	0.002
Rpn1	3L	1.97	0.002
CG4836	3R	1.78	0.003

correlated (Akashi 1996; Eyre-Walker 1996; Comeron et al. 1999; Duret and Mouchiroud 1999; Marais and Duret 2001; Singh et al. 2005a), it is therefore also expected that protein length and the selection parameter should be negatively correlated. For genes with $S > 0$, protein length and the selection parameter estimate are indeed significantly negative correlated (Kendall's $\tau = -0.54$, $P << 0.0001$). Partial correlation analyses reveal that the associations among S, protein length, and codon bias are independent (data not shown).

Intragenomic Comparisons of Selection Parameter Estimates

The strength of selection on synonymous sites in *D. sechellia* is estimated to be quite weak. Median values of S_s were -5.37 and 2.04 for $S_s < 0$ and $S_s > 0$, respectively. Because the estimates of $|S_s|$ and the statistical significance of the rejection of the null model (q value) are significantly correlated (Kendall's $\tau = -0.71$, $P < 0.0001$ and Kendall's $\tau = -0.66$, $P < 0.0001$ for all genes with $S_s > 0$ and all genes with $S_s < 0$, respectively), the estimate of the median strength of selection in the 5% FDR gene set may be biased toward stronger selection parameter estimates. In addition, it is difficult to compare the strengths of selection between the 2 classes of genes ($S_s > 0$ and $S_s < 0$) given that there are only 8 genes in *D. sechellia* with $S_s < 0$.

The lineage-specific estimates of S_s in this gene-by-gene analysis facilitate comparing the nature of selection between the X and the autosomes within *D. sechellia*. There are no significant differences between the mean estimates of the selection parameter of X-linked versus autosomal genes for genes with $S_s > 0$ ($P = 0.37$, Mann–Whitney U test). We can also examine whether autosomal or X-linked genes are over- or underrepresented by comparing the ratio of autosomal and X-linked genes in the full set to their ratio in the set of genes that show evidence of selection on synonymous sites. In *D. sechellia*, the ratio of X-linked genes to autosomal genes is significantly higher in the subset of genes with evidence of selection when $S_s > 0$ ($P = 0.01$, G-test), which is suggestive of either an excess of X-linked genes or a relative dearth of autosomal genes in this subset.

To gain more insight into the types of genes experiencing selection toward preferred codons in *D. sechellia*, we assessed whether certain functional categories of genes had systematically increased strengths of selection on synonymous sites. Only one GO term, "cytoplasm organization and biogenesis," had significantly higher median estimates of S_s than expected ($P = 0.02$, see Methods). We also assessed whether there were any GO terms that were significantly overrepresented in the 5% FDR gene set (see Methods). The cellular component terms "kinetochore" and "integral to membrane" as well as the biological process term "G-protein–coupled receptor protein signaling pathway" are significantly enriched in 5% FDR set ($P \leq 0.042$, all comparisons). The enrichment of genes associated with the "integral to membrane" term may be at least partly driven by genes also associated with the term "G-protein–coupled receptor protein signaling pathway," as of the 27 genes associated with the former, 11 are also associated with the latter.

Finally, we specifically examined patterns of selection at synonymous sites at genes on the likely nonrecombining dot chromosome. Of the 38 genes dot chromosome genes with greater than 5 synonymous substitutions in *D. melanogaster*, 3 of them show statistically significant negative estimates of the selection parameter, although none of these genes survive correction for false discovery. In *D. sechellia*, 9 of 34 dot chromosome genes with more than 5 synonymous substitutions showed evidence in support of selection toward unpreferred codons, and one of these genes appears in the 5% FDR set.

Discussion

The molecular evolutionary process is composed of 2 distinct stages: the mutational generation of novel mutations and the subsequent population process through which these mutations are fixed or lost from the population. The maximum likelihood framework employed here allows for the separation of these 2 phases and estimates not only the relative rates of single nucleotide mutations but also selection parameters in a lineage-specific fashion. As a consequence, this method facilitates investigation of mutation rate variation between species as well as examining whether the nature and strength of selection on synonymous sites varies interspecifically.

Mutational Spectra and Base Composition

Previous inferences of the relative rates of single nucleotide substitutions have relied on nonfunctional sequences, such as pseudogenes and transposable elements. Because these types of sequences are thought to evolve in the absence of selective constraint, the background substitutional patterns revealed by these analyses may primarily reflect underlying mutational patterns. Such studies, which have been limited to mammals and *Drosophila*, have revealed significant intragenomic and interspecific variation in substitution profiles (Petrov and Hartl 1999; Singh et al. 2004, 2006).

The lineage-specific estimation of the relative rates of each of the 12 single nucleotide mutations allows for the investigation of interspecific variation in mutation rates. At a coarse scale, the mutational profiles from *D. melanogaster* and *D. sechellia*, presented in figure 1, bear striking resemblance to the previously documented substitutional profiles from other known *Drosophila* species (Petrov and Hartl 1999; Singh et al. 2004, 2006). One hallmark of the substitutional profile in *Drosophila* is the increased rate of C→T and G→A substitutions; this appears to be driven by the mutational process given that in both X-linked and autosomal genes in both *D. melanogaster* and *D. sechellia*, these transitions occur with the highest frequency. When coupled with previous reports from species as distantly related as *Drosophila virilis* (Petrov and Hartl 1999) and *D. saltans* (Singh et al. 2006), these results are consistent with this elevation being a phenomenon that

is general to all *Drosophila* species. However, unlike previous reports from other *Drosophila* species wherein the rates of A→G and T→C substitutions are similar to that of transversions (Petrov and Hartl 1999; Singh et al. 2004, 2006), the mutation rates of these transitions on both the X and the autosomes appear to have elevated rates in *D. sechellia*. Thus, although there do appear to be features of the mutational profile that are consistent across *Drosophila*, there also appear to be species-specific variations in the mutational spectrum.

While the relative rates of the 4 transitions are those that vary most dramatically between the 2 species, the remaining 8 single nucleotide mutations vary as well. To determine the overall impact of the differences in the relative rates of single nucleotide mutations between species, we estimated equilibrium GC content (π_{GC}) for X-linked and autosomal genes of each species. The π_{GC} is consistently higher in *D. sechellia* than in *D. melanogaster* (fig. 2), which is to be expected given the relative increase in GC-enriching transitions and the relative decrease of the AT-enriching transitions (fig. 1). Consistent with these differences in π_{GC}, comparing the mutation rate estimates of autosomal or X-linked genes between species within a likelihood framework reveals significant differences. This provides statistical support for a shift in mutational patterns in a brief period of evolutionary time, confirming previous observations and inferences (Takano-Shimizu 1999, 2001; Kern and Begun 2005; Akashi et al. 2006) and adds support to the emerging view that mutation rates are relatively labile on an evolutionary timescale.

The mutation rate estimation also facilitates investigating intragenomic variation in mutation rates. For instance, given that the X chromosome spends two-thirds of its evolutionary history in females and one-third of its evolutionary history in males, whereas the autosomes spend half of their evolutionary history in each sex, differences in the mutational profiles of X-linked versus autosomal genes may be consistent with sex-specific mutational biases. In both species, π_{GC} of autosomal or X-linked genes differ by approximately 3–4%. Given our methodology, we cannot currently assess the statistical significance of intragenomic variation in mutation rates, and this is a direction of future research. The magnitude of this intragenomic variation is markedly lower than the interspecific differences in π_{GC}, which are on the order of 7–14%. In addition, this order of magnitude of intragenomic variation of 3–4% is consistent with variation in π_{GC} along a recombination gradient on the autosomes of the *D. melanogaster* genome (Singh et al. 2004). Of particular importance is the observation that in *D. melanogaster* π_{GC} is decreased on the X chromosome, whereas the opposite is true for *D. sechellia*. If these differences in mutational patterns between the X and the autosomes truly arise from sex-specific mutational biases, this may suggest that sex-specific mutational biases are in fact lineage specific.

Finally, we can use the mutation parameter estimates to examine potential nonequilibrium base composition in coding sequences. Specifically, estimates of π_{GC} can be compared with the observed GC content at 4-fold degenerate sites in X-linked and autosomal genes. Observed GC content is markedly greater than expectation for both autosomal and X-linked genes in both *D. melanogaster* and *D. sechellia*, although the magnitude of the departure is more pronounced in *D. melanogaster* (fig. 2). This may in part be explained by selection on synonymous sites; given that most preferred codons in *Drosophila* are G or C ending, selection for increased codon bias will result in an increase in GC content at 4-fold degenerate sites. However, that a departure from expected GC content has been documented previously in noncoding sequences on the autosomes of the *D. melanogaster* genome (Singh et al. 2004) may suggest that the nonequilibrium base composition in coding sequences may not entirely be due to selection on genic function and may instead reflect the difficulty in achieving base composition equilibrium given lineage-specific shifts in mutational patterns. It is also possible that the departure from equilibrium in noncoding sequences is affected by selection on these sequences as well. As a consequence, although base frequencies in both *D. melanogaster* and *D. sechellia* are not at their equilibrium values, determining the ultimate causes of the increased GC content relative to expectation is challenging. Assessing the impact of nonequilibrium base composition in extant species, presumed nonequilibrium ancestral base composition, as well as recent lineage-specific shifts in mutational patterns on the estimation of the strength and direction of selection at synonymous sites is best addressed by a simulation approach and will be a topic of future investigation.

Genomic Correlates of Selection Parameter Estimates

The analysis of genes in the 5% FDR set based on the test of H_0: $S = 0$ allows for the characterization of the relationships between genomic features for precisely the subset of genes with evidence for selection at synonymous sites and for the identification of genic parameters that are associated with such selection. It is important to note that because polymorphic sites within species might be erroneously considered fixed differences between species, our use of a single sequence from each species underestimates the strength of selection for positive selection parameter estimates and overestimates negative selection parameter estimates. This is principally due in a bias in the distribution of polymorphic variants toward unpreferred mutations. This bias may differentially affect the X chromosome and the autosomes in both species as polymorphism on the X chromosome, at least in *D. melanogaster*, is reduced compared with polymorphism on the autosomes (e.g., Andolfatto 2001) and may similarly pose less of a challenge in *D. sechellia*, where standing levels of variation are reduced relative to *D. melanogaster* (Kliman et al. 2000; Morton et al. 2004). However, previous analysis of the impact of conflating polymorphism and divergence on selection parameter estimation suggests that the magnitudes of the increases/decreases in S are minor at best (Nielsen et al. 2007).

Drosophila sechellia: Positive Selection Parameter Estimates

Genes with positive estimates of S appear to be experiencing selection consistent with a traditional model of

selection for increased translational efficiency/accuracy. Because the evolution of codon bias has been extensively studied in this system, we can formulate specific hypotheses to test using our data. For instance, previous reports suggest selection at synonymous sites is weak (Akashi 1995, 1997; McVean and Vieira 2001; Singh et al. 2004). Consistent with these findings, our data suggest that the magnitude of the selection coefficient on codon bias is quite small, with a median value of $S = 2.04$. If N_e is 1×10^6, then the selective advantage (s) of individual alleles is on the order of 1×10^{-6}. Importantly, this method does not account for interference among selected mutations that may have a stronger, yet unknown, effect on the selection parameter estimates. Thus, although these genes do show evidence of selection on synonymous sites that is consistent with selection for increased codon bias, the magnitude of the effect is quite small.

If selection pressure for codon bias is conserved on an evolutionary timescale, then we expect genes with high codon bias to show evidence of strongest selection toward preferred codons. This is precisely what these data suggest as increases in the selection parameter estimate are associated with increases in codon usage bias in *D. sechellia*. Additionally, previous reports have shown that longer sequences generally have lower codon bias (Akashi 1996; Eyre-Walker 1996; Comeron et al. 1999; Duret and Mouchiroud 1999; Marais and Duret 2001; Singh et al. 2005a), which may be due to either Hill–Robertson interference or increased selective benefits of codon bias in shorter versus longer genes. If longer proteins are under weaker selection for preferred codons, we expect a negative correlation between protein length and the selection parameter estimate, which is recovered as well.

Drosophila sechellia: Negative Selection Parameter Estimates

Previous application of this maximum likelihood framework to a set of 18 genes revealed one gene, *Notch*, which showed a significantly negative selection parameter estimate in *D. melanogaster*. The application of this framework at the genomic scale facilitates investigating whether there are other genes that, like *Notch*, have selection parameter estimates that are significantly less than 0. These genes, while likely under selection at synonymous sites, are subject to selection on a feature other than increased codon bias in the traditional framework. Identification of such genes thus provides interesting avenues of future research as investigation of the molecular evolution of these genes may reveal novel selective pressures on synonymous sites.

Negative selection parameter estimates correspond to selection in favor of unpreferred codons, the causes of which are not immediately obvious. First, it is possible that the negative selection parameter estimates are an artifact of a potential misspecification of the mutation rate matrix. If significant heterogeneity in mutational patterns exists within autosomal and X-linked genes, which we have failed to capture by estimating a single mutation profile for each of these chromosome sets, then our estimates of selection may not be entirely accurate. Depending on the nature and the direction of the variations in the rates of the single nucleotide mutations, this could potentially result in spuriously negative selection parameter estimates. It is important to note that a misspecification of the mutation rate matrix is not sufficient to explain the negative estimate of S for *Notch* in *D. melanogaster* as this result is robust to the specification of the mutation matrix based on introns within this gene (Nielsen et al. 2007).

An alternative hypothesis is that genes with negative selection parameter estimates employ a different set of optimal codons than are currently defined for *D. melanogaster*. If genes with $S < 0$ were expressed in particular tissues, or only at particular stages during development and if there were tissue- or developmental stage-specific tRNA pools, it is theoretically possible that another set of optimal codons would be appropriate for these genes.

Additionally, preferred codons in *Drosophila* are GC-biased, and unpreferred codons are largely AT-biased. As a consequence, selection toward unpreferred codons is roughly analogous to selection in favor of increased AT content. If there is selection on base composition of coding sequences in *Drosophila*, genes with negative estimates of the selection parameter may correspond to those genes with selective pressure to increase AT content.

Further, genes with negative parameter estimates may correspond to genes for which reduced codon bias is selectively beneficial. Levels of codon bias affect rates and accuracy of protein translation and have also been implicated in affecting levels of active protein (Carlini and Stephan 2003). It may be that recent changes in selective pressures have resulted in a selective benefit to reduced rates of translation and/or reduced levels of active protein for some genes.

Lastly, molecular evolution at synonymous sites may reflect selection for features beyond translational accuracy/efficiency. Indeed, mutations at synonymous sites can impact protein folding (Kimchi-Sarfaty et al. 2007), and it is similarly possible that synonymous site changes can affect other features such as mRNA stability. Thus, genes with selection parameter estimate significantly less than zero may correspond to genes under selection relating to aspects of mRNA stability or protein folding.

Discussion remains open with respect to the ultimate causes of a negative selection parameter estimate, and given the relative dearth of genes of this type in both *D. melanogaster* and *D. sechellia*, it is difficult to determine whether or not there are defining characteristics of this gene set. In general, however, selection on synonymous sites in this gene class also appears to be weak, with a median estimate of $S = -5.37$ in *D. sechellia*.

Drosophila melanogaster

Only 3 genes in *D. melanogaster* show strong evidence for selection on synonymous sites. Of these, 2 have positive parameter estimates and 1 has a negative selection parameter estimate. This latter gene, *Notch*, was previously identified as evolving in a manner consistent with selection in favor of unpreferred codons (Bauer DuMont et al. 2004; Nielsen et al. 2007). This is particularly intriguing given

that in *D. sechellia* (table 3) and *D. simulans* (Nielsen et al. 2007), *Notch* is evolving in a manner consistent with selection in favor of preferred codons. The paucity of genes in this species with evidence for selection may indeed reflect a relaxation of selection on synonymous sites in this lineage, which has been suggested previously (Akashi 1995, 1996; McVean and Charlesworth 1999; Bauer DuMont et al. 2004; Akashi et al. 2006; Nielsen et al. 2007). However, this may also result from a lack of sufficient power to detect selection in *D. melanogaster*. One possibility is that the chi-squared approximation, which we use to assess the significance of the likelihood ratio, is not accurate in the tail of the distribution. It is not immediately obvious why this confounding factor would more greatly reduce power in *D. melanogaster* and not *D. sechellia*, but the observation that the correlations between S and other genic features of interest reported here in *D. sechellia* are also found in *D. melanogaster* (data not shown) suggests that there may indeed be (or has been until recently) selection on synonymous sites in this lineage as well.

It may also be the case that is more difficult to reject the null model that $S = 0$ for genes with negative selection parameter estimates, and this is an avenue of future research. Were this the case, this would more greatly affect *D. melanogaster* than in *D. sechellia* because a larger proportion of genes show $S_m < 0$ than $S_s < 0$. This could ultimately shift the distribution of P values from the likelihood ratio test to the right in *D. melanogaster* relative to the distribution of P values from *D. sechellia*, which would result in fewer loci surviving a correction for multiple tests in the former. Further investigation of the nature and strength of selection on synonymous sites in *D. melanogaster* is thus warranted.

Selection Parameter Estimates: Intragenomic Comparisons

The estimation of lineage-specific selection parameters in a gene-by-gene manner allows for the direct comparison of the nature and strength of selection within *D. sechellia*. In particular, we can ask whether the autosomes and the X chromosome differ with respect to patterns of selection on synonymous sites. There is an overrepresentation of X-linked genes (or a dearth of autosomal genes) in the subset of genes with evidence for selection in *D. sechellia*. This may suggest that more genes on the X chromosome are subject to selection on synonymous sites or this excess of X-linked genes in the selected subset may result from an increased efficacy of selection on the X, as the estimated strength of selection on synonymous sites was not found to be greater on the X chromosome. This is consistent with previous observations of increased codon bias in X-linked genes in *D. melanogaster* (Comeron et al. 1999; Hambuch and Parsch 2005; Singh et al. 2005b), *D. pseudoobscura* (Singh et al. 2005b), as well as other *Drosophila* species including *D. sechellia* (Drosophila 12 Genomes Consortium, forthcoming).

To gain additional insight into the types of genes experiencing significant selection at synonymous sites, we investigated whether certain types of genes had characteristically increased estimates of the strength of selection toward preferred codons in *D. sechellia*. To do so, we used a custom GO-SLIM database with 105 GO terms and tested whether certain terms had significantly higher median estimates of S than expected given the distribution of the estimates of S in the set of genes that significantly reject the null model. Only one GO term, cytoplasm organization and biogenesis was associated with stronger selection for $S > 0$ in *D. sechellia*. Thus, genes involved in the assembly and arrangement of the cytoplasm and its components may be under the strongest selection for preferred codons.

Three additional GO terms are significantly overrepresented in the 5% FDR set: kinetochore, integral to membrane, and G-protein–coupled receptor protein signaling. These results suggest that genes involved in the formation of the kinetochore complex, transmembrane proteins, and genes involved in the signaling pathway generated as a consequence of a G-protein–coupled receptor binding to its physiological ligand may be most likely to be under selection for increased codon bias in *D. sechellia*.

Conclusions

Our analyses of coding sequence evolution in *D. melanogaster* and *D. sechellia* are suggestive of both intra- and interspecific differences in mutational patterns as well as patterns of selection at synonymous sites. In particular, these data hint at the possibility of sex-specific mutational biases that may vary among lineages and are consistent with variations in mutational profiles between species as well. With respect to patterns of evolution at synonymous sites, in both *D. melanogaster* and *D. sechellia*, we find evidence in support of individual genes evolving under selection both in favor of and against preferred codons. At a genomic scale, we find very little evidence in support for selection on synonymous sites in *D. melanogaster* as only 3 genes have strong evidence for selection. In contrast, almost 80 times as many genes show strong evidence for selection at synonymous sites in *D. sechellia*. These data thus highlight not only differences in the relative rates of single nucleotide mutations within and between species but also variations in both the nature and strength of selection on synonymous sites between species.

It is difficult to speculate as to the ultimate cause of the variation in patterns of mutation and selection between *D. melanogaster* and its close relative *D. sechellia*. Clearly, differences in the natural history characteristics between these species (e.g., the cosmopolitan versus island-endemic distribution of *D. melanogaster* vs. *D. sechellia*) are likely to contribute to some degree. Other factors such as demographic history of these species may also play a role. However, that our analyses reveal marked changes in the nature of selection at synonymous sites between *D. melanogaster* and its sister species indicates that patterns of mutation and selection can and do change rapidly on an evolutionary timescale. The consequences of these mutational and selective shifts as well as the underlying causes of these changes merit further investigation.

Acknowledgments

The authors gratefully acknowledge M. Hamblin, R. Durrett, D. Schmidt, and A. Wong for comments on this

manuscript and members of the Aquadro lab for fruitful discussion regarding this project. We are especially indebted to A. Larracuente and T. Sackton for their contributions and their marked assistance with respect to the GO analysis and masking of multispecies alignments, respectively. Comments from 2 anonymous reviewers improved the quality of this manuscript. Partial support for this work was provided by National Institutes of Health grant GM36431 to C.F.A. and National Science Foundation grant DMS-0201037 to R. Durrett, C.F. Aquadro and R. Nielsen.

Literature Cited

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.

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.

Akashi H, Kliman RM, Eyre-Walker A. 1998. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Genetica. 102–103:49–60.

Akashi H, Ko WY, Piao SF, John A, Goel P, Lin CF, Vitins AP. 2006. Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics. 172:1711–1726.

Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. Genetics. 146:295–307.

Anderson CL, Carew EA, Powell JR. 1993. Evolution of the Adh locus in the *Drosophila willistoni* group: the loss of an intron, and shift in codon usage. Mol Biol Evol. 10:605–618.

Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol. 18:279–290.

Bauer DuMont V, Fay JC, Calabrese PP, Aquadro CF. 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.

Bulmer M. 1988. Are codon usage patterns in unicellular organisms determined by selection mutation balance? J Evol Biol. 1:15–26.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129:897–908.

Carlini DB, Stephan W. 2003. *In vivo* introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. Genetics. 163:239–243.

Castillo-Davis CI, Hartl DL. 2003. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 19:891–892.

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

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

Cutter AD, Payseur BA, Salcedo T, et al. (12 co-authors). 2003. Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res. 13:2651–2657.

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

Drosophila 12 Genomes Consortium. Forthcoming. Evolution of Genes and Genomes on the Drosophila Phylogeny. Nature.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol. 13:864–872.

Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol Biol Evol. 11:725–736.

Hambuch TM, Parsch J. 2005. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. Genetics. 170:1691–1700.

Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics. 160:595–608.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli system. J Mol Biol. 151:389–409.

Kern AD, Begun DJ. 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.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the *MDR1* gene changes substrate specificity. Science. 315:525–528.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics. 156:1913–1931.

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

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. J Mol Evol. 52:275–280.

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

Marais G, Mouchiroud D, Duret L. 2003. Neutral effect of recombination on base composition in *Drosophila*. Genet Res. 81:79–87.

McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res. 74:145–158.

McVean GAT, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics. 157:245–257.

Morton RA, Choudhary M, Cariou ML, Singh RS. 2004. A reanalysis of protein polymorphism in *Drosophila melanogaster*, *D. simulans*, *D. sechellia* and *D. mauritiana*: effects of population size and selection. Genetica. 120: 101–114.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.

Nielsen R, Bauer DuMont V, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol Biol Evol. 24:228–235.

Petrov DA, Hartl DL. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc Natl Acad Sci USA. 96:1475–1479.

Powell JR, Sezzi E, Moriyama EN, Gleason JM, Caccone A. 2003. Analysis of a shift in codon usage in *Drosophila*. J Mol Evol. 57:S214–S225.

Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. Genome Res. 11: 230–239.

Richards S, Liu Y, Bettencourt BR, et al. (52 co-authors). 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 1999a. Molecular evolution and phylogeny of the *Drosophila saltans* species group inferred from the *Xdh* gene. Mol Phylogen Evol. 13:110–121.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 1999b. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics. 153: 339–350.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2000a. Evidence for a high ancestral GC content in *Drosophila*. Mol Biol Evol. 17:1710–1717.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2000b. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. J Mol Evol. 50:1–10.

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

Singh ND, Arndt PF, Petrov DA. 2004. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics. 169:709–722.

Singh ND, Arndt PF, Petrov DA. 2006. Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. BMC Biol. 4:doi: 10.1186/1741-7007-1184-1137

Singh ND, Davis JC, Petrov DA. 2005a. Codon bias and noncoding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. J Mol Evol. 61:315–324.

Singh ND, Davis JC, Petrov DA. 2005b. X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. Genetics. 171:145–155.

Storey JD. 2002. A direct approach to false discovery rates. J Roy Stat Soc Ser B. 64:479–498.

Takano-Shimizu T. 1999. Local recombination and mutation effects on molecular evolution in *Drosophila*. Genetics. 153:1285–1296.

Takano-Shimizu T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. Mol Biol Evol. 18:606–619.

Vinogradov AE. 2001. Intron length and codon usage. J Mol Evol. 52:2–5.

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

Spencer Muse, Associate Editor

Accepted September 7, 2007