

# Strong Evidence for Lineage and Sequence Specificity of Substitution Rates and Patterns in *Drosophila*

Nadia D. Singh,\* Peter F. Arndt,† Andrew G. Clark,\* and Charles F. Aquadro\*

\*Department of Molecular Biology and Genetics, Cornell University; and †Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin

Rates of single nucleotide substitution in *Drosophila* are highly variable within the genome, and several examples illustrate that evolutionary rates differ among *Drosophila* species as well. Here, we use a maximum likelihood method to quantify lineage-specific substitutional patterns and apply this method to 4-fold degenerate synonymous sites and introns from more than 8,000 genes aligned in the *Drosophila melanogaster* group. We find that within species, different classes of sequence evolve at different rates, with long introns evolving most slowly and short introns evolving most rapidly. Relative rates of individual single nucleotide substitutions vary ~3-fold among lineages, yielding patterns of substitution that are comparatively less GC-biased in the *melanogaster* species complex relative to *Drosophila yakuba* and *Drosophila erecta*. These results are consistent with a model coupling a mutational shift toward reduced GC content, or a shift in mutation–selection balance, in the *D. melanogaster* species complex, with variation in selective constraint among different classes of DNA sequence. Finally, base composition of coding and intronic sequences is not at equilibrium with respect to substitutional patterns, which primarily reflects the slow rate of the substitutional process. These results thus support the view that mutational and/or selective processes are labile on an evolutionary timescale and that if the process is indeed selection driven, then the distribution of selective constraint is variable across the genome.

## Introduction

Rates and patterns of molecular evolution are governed by the dynamic interplay among mutation, random genetic drift, biased gene conversion (BGC), and natural selection. As a consequence, population genetic parameters that modulate the effects of these forces can lead to variation in rates of single nucleotide substitution within genomes as well as between genomes. Effective population size, coefficients of selection, recombination rate, dominance coefficients, and chromosomal location, among other factors, are all likely to contribute to heterogeneity in intra and intergenomic heterogeneity in substitution rate.

There is a wealth of literature documenting heterogeneity in rates and patterns of evolution within species in *Drosophila*. For instance, several functional classes of genes are known to evolve particularly rapidly, such as immune-related genes (e.g., Schlenke and Begun 2003; Sackton et al. 2007; for a review see Lazzaro 2008), sex- and reproduction-related genes (e.g., Begun et al. 2000; Swanson et al. 2001; for a review see Swanson and Vacquier 2002; Haerty et al. 2007), and genes with sex-biased expression patterns (e.g., Zhang et al. 2004; Proschel et al. 2006; for a review see Ellegren and Parsch 2007; Zhang et al. 2007). Factors influencing the strength of purifying selection on proteins as well as features affecting the degree to which individual proteins are subject to positive selection also contribute to variation in evolutionary rate among proteins. These factors include expression level, developmental timing of expression, recombination rate, gene essentiality, and gene structure features such as intron presence/absence, intron length protein length, and contact density (e.g., Betancourt and Presgraves 2002; Marais et al. 2004, 2005; Davis et al. 2005; Lemos et al. 2005; Zhang and Parsch 2005; Larracuentte et al. 2008; Zhou et al.

2008). Chromosomal linkage also has the potential to significantly affect rates of evolution in X-linked versus autosomal sequences (Avery 1984; Charlesworth et al. 1987; Betancourt et al. 2004). Although in *Drosophila*, while some evidence suggests X-linked genes evolve more rapidly than autosomal genes (Betancourt et al. 2002; Thornton and Long 2002; Counterman et al. 2004; Begun et al. 2007; Baines et al. 2008), this does not appear to be a universal trend (Thornton et al. 2006; Singh et al. 2008).

Noncoding sequences in *Drosophila* have also been shown to have variable evolutionary rates and appear generally more highly constrained than synonymous sites (Bergman and Kreitman 2001; Halligan et al. 2004; Kohn et al. 2004; Andolfatto 2005; Haddrill et al. 2005, 2008; Kern and Begun 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006; Haddrill and Charlesworth 2008). Patterns of constraint and divergence in intronic and intergenic sequences and in untranslated regions (UTRs) appear variable as well, although the precise distribution of selective constraint across noncoding sequences remains somewhat unclear. Although initial studies indicated that patterns of constraint were similar between conserved intergenic and intronic regions (Bergman and Kreitman 2001), later studies suggested that introns and intergenic regions do, in fact, differ with respect to degree of selective constraint (Halligan et al. 2004; Andolfatto 2005; Bachtrog and Andolfatto 2006), as do UTRs (Andolfatto 2005; Haddrill et al. 2008). More recent evidence further suggests that patterns of divergence in noncoding sequence are likely functionally related to sequence length, with short introns and intergenic sequences evolving more rapidly than longer noncoding sequences (Haddrill et al. 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006), which does not appear to be due to differences in underlying mutation rate (Wang et al. 2007).

In addition to varying within genomes, evolutionary rates have been shown to vary among species in *Drosophila*. It is easy to intuit how at least some protein-coding genes could evolve in a lineage-specific fashion, given the morphological, behavioral, and ecological diversity encompassed by the genus *Drosophila*. Indeed, recent comparative

Key words: substitutional patterns, *Drosophila*, stationary GC content, selective constraint.

E-mail: nds25@cornell.edu.

*Mol. Biol. Evol.* 26(7):1591–1605. 2009

doi:10.1093/molbev/msp071

Advance Access publication April 7, 2009

genomic efforts have identified several genes and gene families that appear to be evolving in a manner that is consistent with lineage- or clade-specific selective pressures (e.g., *Drosophila* 12 Genomes Consortium 2007; McBride 2007; McBride and Arguello 2007). Even closely related species such as *Drosophila melanogaster* and *Drosophila simulans* often show markedly different rates of evolution in orthologous genes (e.g., Eanes 1994; Akashi 1996; Takano 1998; Nielsen et al. 2007), suggesting that factors contributing to evolutionary rate heterogeneity can be extremely labile on an evolutionary timescale.

Beyond heterogeneity in evolutionary rate within and among *Drosophila* species, patterns of coding and noncoding sequence evolution are known to vary significantly as well. For instance, patterns of codon usage in protein-coding sequences have been shown to vary among *Drosophila* genomes (e.g., Vicario et al. 2007), most notably in the *Drosophila saltans* and *Drosophila willistoni* lineages (Anderson et al. 1993; Rodriguez-Trelles et al. 1999, 2000a, 2000b; Powell et al. 2003; Heger and Ponting 2007; Vicario et al. 2007). Like protein evolutionary rate, intragenomic variation in codon bias has been associated with a multitude of factors including recombination rate, protein length, expression level, gene structure, gene density, and chromosomal linkage (Kliman and Hey 1993; Akashi 1996; Eyre-Walker 1996; Comeron et al. 1999; Duret and Mouchiroud 1999; Comeron and Kreitman 2000; Marais and Duret 2001; Marais et al. 2001, 2003; Vinogradov 2001; Hey and Kliman 2002; Hambuch and Parsch 2005; Singh et al. 2005, 2008). The degree of GC bias in segregating polymorphisms and single nucleotide substitution varies in noncoding sequences in *Drosophila* as well, both within and between genomes (Takano-Shimizu 2001; Singh et al. 2004, 2006; Kern and Begun 2005; Galtier et al. 2006; Ko et al. 2006; Ometto et al. 2006; Wang et al. 2007; Haddrill and Charlesworth 2008).

The bulk of previous literature on variation in rates and patterns of molecular evolution in *Drosophila* have been, by necessity, limited in scale or in scope and focused on subsets of genes and/or individual or pairs of lineages. The recent sequencing of additional *Drosophila* species (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007) facilitates systematically and quantitatively investigating these questions in a number of species at a genomic scale. Here, we statistically evaluate and quantify heterogeneity in rates and patterns of single nucleotide substitution at a genomic scale in a lineage-specific fashion in *Drosophila*. We applied a maximum likelihood model to 4-fold degenerate synonymous sites and introns from over 8,000 genes aligned in the *D. melanogaster* group, which facilitated estimating substitution rates on individual branches of the phylogeny. Our results indicate statistically significant interspecific differences in patterns of single nucleotide substitution in both coding and intronic sequences, with relative rates of single nucleotide substitution varying up to 3-fold among species. Moreover, these data suggest that there may have been a shift in mutational patterns or a shift in mutation–selection balance toward increased AT content in the *D. melanogaster* species complex. In addition, these data unequivocally show that different classes of sequence within the genome evolve in significantly different ways, with X versus autosomal linkage

playing a significant role. Although both of these patterns have been reported previously in isolated lineages (particularly *D. melanogaster* and *D. simulans*), our analysis indicates that this heterogeneity in evolutionary rate among sequence types is characteristic of the entire *melanogaster* group.

Our analysis also facilitates estimating the magnitude of variation in evolutionary rate among sequence classes. Specifically, our data suggest that short introns evolve at tantalizingly close to the neutral rate, whereas long introns appear to be evolving up to 50% less rapidly than 4-fold degenerate synonymous sites. Finally, although previous work has suggested nonequilibrium base composition at a smaller scale (Singh et al. 2004, 2007; Kern and Begun 2005; Akashi et al. 2006), our analysis indicates that base composition at synonymous sites and in intronic sequences is not at equilibrium with respect specifically to patterns of single nucleotide substitution throughout the phylogeny. We quantify this departure from equilibrium, and note that the extent to which base composition deviates from equilibrium varies in a lineage-specific fashion. We argue that this is primarily due to the very slow timescale at which base composition equilibrates following changes in substitution patterns.

These results are likely to have profound implications for myriad molecular evolutionary studies. For instance, most traditional phylogenetic models assume that all branches on the phylogeny can be characterized by a single nucleotide transition matrix and that base composition is at equilibrium with respect to this matrix, and our results clearly show that both of these assumptions are violated in *Drosophila*. In addition, the nonequilibrium nature of base composition in *Drosophila* may compromise our ability to make population genetic inferences, as many population genetic models assume base composition equilibrium. Moreover, this analysis sheds light on the distribution and relative strength of selective constraint throughout the genome in *Drosophila*, which will be invaluable as neutral evolutionary models continue to develop.

## Methods

### Sequence Data

There are six sequenced species in the *D. melanogaster* group: *D. melanogaster*, *D. simulans*, *Drosophila sechellia*, *D. erecta*, *D. yakuba*, and *Drosophila ananassae*. We retrieved masked coding sequence alignments for all single-copy orthologs in these six species from Flybase ([ftp://ftp.flybase.net/12\\_species\\_analysis](ftp://ftp.flybase.net/12_species_analysis)) (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007). These alignments correspond to the longest *D. melanogaster* transcript for each gene. Fourfold degenerate synonymous sites (from amino acids conserved across the *melanogaster* group) were extracted from a total of 8,563 genes and concatenated together, treating the X chromosome and the autosomes separately, yielding 1.3-Mb autosomal sequence and 0.18-Mb X-linked sequence aligned across the *melanogaster* group. Introns from these genes were retrieved from the University of California, Santa Cruz genome browser (<http://genome-test.cse.ucsc.edu/cgi-bin/hgGateway>). We provided a list of

FlyBase gene numbers, and all intronic sequences associated with these gene models were extracted using the Custom Track feature. We restricted ourselves to those alignments for which sequence was available for all six species in the *D. melanogaster* group. Coding sequences within introns were removed, and the 10 bp adjacent to intron–exon boundaries (based on *D. melanogaster* gene models) were stripped from the alignments to remove potentially functionally constrained sites. These filtered introns were then separated into short ( $\leq 65$  bp) and long introns ( $> 65$  bp), following a previous suggestion (Halligan and Keightley 2006), and introns within each size class were concatenated together for the X chromosome and autosomes separately. This procedure resulted in 11.0 and 0.28 Mb of aligned autosomal long and short intron sequences, respectively, and, respectively, 1.5 Mb and 23.9 kb for the X chromosome. Classification of X-linked versus autosomal location was based on the genomic location in *D. melanogaster*. Although there are genes that have moved among chromosomes within the *melanogaster* group, the rate of interchromosomal gene traffic in *Drosophila* appears to be quite low (Ranz et al. 2001; Richards et al. 2005; Bhutkar et al. 2007), and we are therefore confident that our results are unlikely to be substantially affected by interchromosomal gene traffic. *Drosophila melanogaster* gene models were used to define sequences as intronic, as well as to infer distances of intronic sites to intron–exon junctions.

### Estimation of Substitution Frequencies

In this work, we make use of a recently developed maximum likelihood method to estimate substitution frequencies from multiple alignments of extant DNA sequence data (Duret and Arndt 2008). The full details of the model have been reported previously (Duret and Arndt 2008), and here we will only briefly summarize key features of this model and describe its application to the current data. With this model, substitution frequencies can be measured independently in all branches of a given phylogeny except for the two branches that connect to the root node (Duret and Arndt 2008). Our formalism allows us to estimate the 12 different substitution frequencies  $r_{\alpha \rightarrow \beta}$  for exchanges of one nucleotide  $\alpha \in \{A, C, G, T\}$  by another  $\beta$ .

We make several simplifying assumptions regarding the single nucleotide substitutional process in this implementation. First, we do not consider the possibility of context-dependent substitutional processes. There is no germline methylation in *Drosophila* (Lyko et al. 2000), which suggests that unlike patterns found in humans and other vertebrates (Arndt and Hwa 2005), CpG substitutions should not occur with increased frequency. This is supported in *D. melanogaster* (Singh et al. 2004) but has yet to be tested explicitly in other species in this genus. There is limited evidence in support of other context-dependent substitutional processes in *D. melanogaster* (Arndt and Hwa 2005; Liu and Li 2008), but the magnitude of the effect appears quite weak as compared with the effect in vertebrates (Arndt and Hwa 2005). As a consequence, we have not allowed for the possibility of context-dependence in our substitution rates estimation, and we do not believe that our results are systematically or substantially affected by this simplification.

Second, we assumed strand complementarity of the nucleotide substitution process. Although there does appear to be evidence in support of strand noncomplementarity of the substitutional process in *Drosophila* (Singh et al. 2004; Nielsen et al. 2007), it is yet unclear what drives this pattern. Moreover, the differences between the rates of complementary substitution are quite small in magnitude (Singh et al. 2004). Although not accounting for strand noncomplementarity may introduce noise into our results, given how minor the effects of strand noncomplementarity appear to be in *D. melanogaster*, we do not believe that this assumption compromises our results.

Third, this model does not take into account rate heterogeneity among sites. Because we had a priori expectations that certain sites might evolve in significantly different ways, we separated X-linked from autosomal sequences, short from long introns and introns from synonymous sites. Within these sequence categories, however, we assume all sites are equivalent with respect to rates of substitution. Finally, we do not account for incomplete lineage sorting or variation in gene trees among sequences, although we specifically test whether our results are robust to treespace (see below).

In our model, the substitution rates are reverse complement symmetric and pairs of rates are equal (e.g.,  $r_{C \rightarrow A} = r_{G \rightarrow T} = r_{C:G \rightarrow A:T}$ ), which yields six substitution frequencies. These rates can be written into a single nucleotide transition matrix

$$Q = \begin{pmatrix} \bullet & r_{C:G \rightarrow A:T} & r_{G:C \rightarrow A:T} & r_{T:A \rightarrow A:T} \\ r_{T:A \rightarrow G:C} & \bullet & r_{G:C \rightarrow C:G} & r_{T:A \rightarrow C:G} \\ r_{T:A \rightarrow C:G} & r_{G:C \rightarrow C:G} & \bullet & r_{T:A \rightarrow G:C} \\ r_{T:A \rightarrow A:T} & r_{G:C \rightarrow A:T} & r_{C:G \rightarrow A:T} & \bullet \end{pmatrix},$$

where the diagonal is constrained by the condition that the elements in one row have to sum to zero, ensuring the conservation of total probability. With this notation the evolution of DNA sequences is governed by the differential equation  $(\partial/\partial t)p = Qp$ , where  $p$  is the vector of nucleotide frequencies  $p = (p_A, p_C, p_G, p_T)^t$ , and  $t$  denotes the vector transposition. We allow for independent sets of substitution frequencies along all branches and use a maximum likelihood framework to estimate these substitution frequencies on each branch given a sequence alignment.

In this article, we take the branch length as the timescale. Consequently, the substitution rates  $r$  measure the frequency of substitutions per nucleotide during the time interval that spans a particular branch in a phylogenetic tree. If this time is known in terms of the physical time, that is, in million years, substitution rates per nucleotide and time could be calculated. However, such information is not always available; in the following, we will therefore mostly analyze quantities, such as the stationary GC content, which are independent of the timescale.

Due to finite amount of sequence data, all estimates of substitution frequencies are subject to stochastic error. We can estimate the amount of uncertainty in the substitution frequencies and nucleotide distributions by bootstrapping the sequence data. The original data are sampled with replacement to generate replicate data sets, and respective substitution frequencies are estimated. The variance in

substitution rate estimates from several bootstrap samples can then serve as an estimator of the variance from the measurement on the original data. Throughout this article, we used 500 bootstrap samples for this procedure.

### Time to Equilibrium Calculations

To investigate the rate of change of GC content, we first notice that the equilibrium nucleotide distribution  $\pi$  is given by  $Q\pi=0$ . For reverse complement symmetric substitutions, the stationary nucleotide distribution is

$$\pi = (1 - f_{\text{stat}}, f_{\text{stat}}, f_{\text{stat}}, 1 - f_{\text{stat}})^t/2,$$

where

$$f_{\text{stat}} = \frac{r_{T:A \rightarrow G:C} + r_{T:A \rightarrow C:G}}{r_{T:A \rightarrow G:C} + r_{T:A \rightarrow C:G} + r_{G:C \rightarrow A:T} + r_{C:G \rightarrow A:T}}$$

is the stationary GC content.

To make statements about the convergence of the GC content to its stationary value, we have to solve the time evolution of the above differential equation, which can be achieved by diagonalizing the  $Q$  matrix. Let  $S$  be the matrix of eigenvectors and  $D = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$  be the diagonal matrix of eigenvalues such that  $QS = SD$ . With this notation, we have  $p(t - t_0) = e^{Qt}p(t_0) = Se^{Dt}S^{-1}p(t_0) = S \text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t})S^{-1}p(t_0)$ .

The matrix  $Q$  has the following first and second eigenvalues:  $\lambda_1=0$ ,  $\lambda_2 = -(r_{T:A \rightarrow G:C} + r_{T:A \rightarrow C:G} + r_{G:C \rightarrow A:T} + r_{C:G \rightarrow A:T})$ . Because in most genomic sequences, the A content is similar to the T content and the C content is close to the G-content (Lobry and Lobry 1999), the initial distribution  $p(t_0)$  has the following property:  $p_A(t_0) \approx p_T(t_0)$  and  $p_C(t_0) \approx p_G(t_0)$ . As a consequence, the third and fourth eigenvalues do not contribute substantially to the time evolution of the GC content. We thus approximate the dynamics by considering only the first two eigenvalues. The dynamics of GC content  $f$  can be approximated by  $f(t) = f_{\text{stat}} + (f_0 - f_{\text{stat}})e^{\lambda_2 t}$  where  $f_0$  is initial GC content. Given this formulation, the GC content half-life, or the time in which the difference between  $f_0$  and  $f_{\text{stat}}$  is reduced by a factor of two, can be simply written as  $t_{1/2} = -\ln(2)/\lambda_2$ . We can estimate time to equilibrium in a similar fashion, but doing so requires defining equilibrium, as strictly speaking, equilibrium will only be reached at  $t = \infty$ . For the purposes of this paper, we define ‘‘equilibrium’’ as within 1% of  $f_{\text{stat}}$  (i.e.,  $f_{t_{\text{eq}}} = f_{\text{stat}} \pm 0.01$ ). We can then estimate

$$t_{\text{eq}} = \frac{-\ln\left(\frac{\text{abs}(f_0 - f_{\text{stat}})}{0.01}\right)}{\lambda_2}.$$

We take the absolute value of the deviation of the current GC content from the stationary one, because the stationary GC content might be approached from above or below.

## Results

### Model Fitting

Because our model was run in a full likelihood framework, we can statistically compare the fit of our data with

various models by using likelihood ratio tests (LRTs) to compare the likelihoods of fully nested models. The distribution of the difference in the log-likelihood values should be approximately  $\chi^2$ , and the number of degrees of freedom corresponds to the difference in the number of free parameters between the nested models. To test for lineage specificity of substitution rates, we implemented a simple model, wherein the nucleotide substitution matrix was constant across the entire phylogeny of the *D. melanogaster* group (supplementary fig. 1, Supplementary Material online), and only branch lengths were allowed to vary. We compared the likelihood of this model with the more highly parameterized model, in which a transition matrix is inferred on each branch of the phylogeny; this full model has 45 additional free parameters as compared with the simple model. We ran both models on the concatenated autosomal data (introns + 4-fold degenerate synonymous sites), the concatenated X chromosome data (introns + 4-fold degenerate synonymous sites), as well as on a combined data set in which X-linked and autosomal sequences were combined. In all cases, accounting for lineage-specific substitution rates dramatically improved the fit of the data to the specified model ( $P < 0.0001$ , all comparisons, LRT).

Given this interspecific variation in substitution rates, we tested for intragenomic differences in rates of single nucleotide substitution. We first examined whether X-linked and autosomal sequences were better characterized by individual transition matrices rather than a single rate matrix for both chromosome sets. We ran the full model on the concatenated autosomal data (introns + 4-fold degenerate synonymous sites) and the concatenated X chromosome data (introns + 4-fold degenerate synonymous sites), as well as on the concatenated X-linked and autosomal sequence. If accounting for interchromosomal heterogeneity in substitution rates substantially improves the fit of the model to the data, then the sum of the log-likelihoods of the two individual models (X chromosome and autosomes, separately) should exceed the log-likelihood of the combined model (X and autosomes concatenated). Indeed, this is precisely what is shown, and the difference in likelihoods is highly statistically significant ( $P < 0.0001$ , LRT).

To further test intragenomic variability in substitution rates, we examined whether 4-fold degenerate sites had significantly different substitutional patterns than introns. We approached this in a parallel fashion to the X versus autosome comparison, in that we compared the log-likelihood of the models on the individual sequence classes (introns and 4-fold degenerate synonymous sites separated) with the combined model (introns and 4-fold degenerate sites concatenated). Separating introns from 4-fold degenerate synonymous sites significantly improved the fit of the data to the specified model ( $P < 0.0001$ , LRT).

Finally, we investigated whether introns of different lengths have significantly different patterns of single nucleotide substitution. We separated X-linked and autosomal introns into two size classes, short ( $\leq 65$  bp) and long ( $> 65$  bp) and ran the full model on these data sets individually. We compared the sum of the log-likelihoods of these two models with the log-likelihood of the model run on all X-linked or autosomal intron sequence concatenated together. For both the X chromosome and the autosomes,

allowing for transition matrices unique to each size class of intron significantly improved the fit of the data to the model ( $P < 0.0001$ , both comparisons, LRT). Additional tests were performed to ensure that the model was behaving as expected and was not suffering from overfitting our data (Supplementary Material online).

### Robustness to Treespace

The *melanogaster* species group consists of six sequenced species: *D. melanogaster*, *D. sechellia* and *D. simulans*, *D. erecta*, *D. yakuba*, and *D. ananassae*, and the *melanogaster* species subgroup includes all of these species except *D. ananassae* (supplementary fig. 1a, Supplementary Material online). The placement of the *melanogaster* species complex (*D. melanogaster*, *D. sechellia*, and *D. simulans*) relative to *D. yakuba* and *D. erecta* is somewhat uncertain, likely due to ancestral polymorphism. Although roughly half of the genes that can be aligned in the *melanogaster* group support a topology in which *D. yakuba* and *D. erecta* are sister species, approximately one-quarter of genes support a topology in which *D. erecta* is the outgroup to the *melanogaster* complex, and the remaining genes support a topology in which *D. yakuba* is the outgroup to the *melanogaster* complex (Pollard et al. 2006; Larracuenta et al. 2008). To test whether our analysis was robust to variation in treespace, we ran the full model on all three tree topologies (supplementary fig. 1, Supplementary Material online) on four different classes of sequence data: 4-fold degenerate synonymous sites on the X and autosomes and intronic sequences on the X and autosomes. In all four data sets, the tree topology in which *D. yakuba* and *D. erecta* are sister species yields a significantly greater likelihood than both alternative topologies ( $P < 0.0001$ , all comparisons, LRT). More importantly, the substitution rate matrices inferred for *D. yakuba* and *D. erecta*, the species most sensitive to alternative placement on the phylogeny, appear robust to variation in treespace (supplementary figs. 2 and 3, Supplementary Material online). Although there may be subtle differences in the estimated transition matrices among the three trees within *D. yakuba* and *D. erecta*, these differences are small on the scale of the differences among species (supplementary fig. 4, Supplementary Material online). Within synonymous sites on the autosomes, for instance, inferred equilibrium GC content varies by 0.5% and 2% among alternative tree topologies within *D. erecta* and *D. yakuba*, respectively, whereas the smallest interspecific difference in equilibrium GC content is at least twice that. We thus used the tree topology in which *D. yakuba* and *D. erecta* are sister species and are confident that the phylogenetic incongruence among different sequences within the genome does not substantially or adversely affect our results.

### Intergenomic Variation in Substitution Rates

The fully parameterized model implemented here allows for maximum likelihood estimation of single nucleotide transition matrices for each branch on the phylogeny with the exception of those connected to the root. Application of this model to coding and intronic sequences in the *melanogaster*

group revealed strong statistical support for the lineage specificity of single nucleotide transition matrices among the five species in the *melanogaster* subgroup (rates of substitution could not be formally quantified in *D. ananassae* because this species is the outgroup to the *melanogaster* subgroup, and this branch is thus connected to the root).

The relative rates of the six pairs complementary of single nucleotide substitutions in each of the six classes of sequence considered here are presented in figure 1. Relative rates are calculated by normalizing the particular single nucleotide substitution rate of interest by the total substitution rate for that sequence class in that particular species. Relative rates of each of the six pairs of complementary nucleotide substitution vary up to 3-fold among species, though to different degrees among sequence classes. Within long intron sequences, substitution rates vary 1.13- to 1.83-fold on the autosomes and 1.07- to 1.82-fold on the X chromosome, and short introns show similar fold variation (1.12- to 1.71-fold and 1.34- to 2.14-fold variation on the autosomes and X chromosome, respectively). Relative rates of substitution at synonymous sites also vary among lineages, with 1.10- to 2.15-fold and 1.19- to 3.04-fold variation on the autosomes and X chromosome, respectively.

The overall substitutional profile appears to be well conserved across species, with no strong transition bias evident in any sequence classes. The ratio of transitions to transversions, which we define as

$$\frac{T_s}{T_v} = \frac{r_{C:G \rightarrow T:A} + r_{T:A \rightarrow C:G}}{r_{C:G \rightarrow A:T} + r_{T:A \rightarrow G:C} + r_{C:G \rightarrow G:C} + r_{T:A \rightarrow A:T}},$$

across sequence type ranges from 1.06 in *D. sechellia* to 1.16 in *D. erecta*, which is consistent with previous reports (Moriyama and Powell 1996; Petrov and Hartl 1999; Wang et al. 2007). However, the relative frequencies of the two transitions vary across species. For instance, in *D. melanogaster* the T:A  $\rightarrow$  C:G transition occurs at a rate similar to each of the transversions, echoing previous findings (Petrov and Hartl 1999; Singh et al. 2004). In all other species, the rate of this transition appears higher (fig. 1).

To investigate the overall impact of this variation in the relative rates of these six pairs of single nucleotide substitutions, we estimated stationary GC content based on the stationary distribution of the transition matrix in the five terminal branches of the *melanogaster* subgroup phylogeny as well as on the internal branches of the *melanogaster* group phylogeny that are not connected to the root. Stationary GC content varies up to 2.4-fold across species (fig. 2, supplementary table 1, Supplementary Material online), with *D. melanogaster* heavily AT-biased (ranging among sequence types from 25% to 35% on the autosomes, 24–31% on the X), for instance, whereas *D. erecta* (41–66% autosome, 45–70% X) and *D. yakuba* (30–62% autosomes, 34–73% X) are comparatively more GC-biased.

### Intragenomic Variation in Rates and Patterns of Substitution: Introns and Synonymous Sites

In addition to intergenomic comparisons, variation in rates and patterns of single nucleotide substitution can be examined intragenomically for each lineage. Specifically,

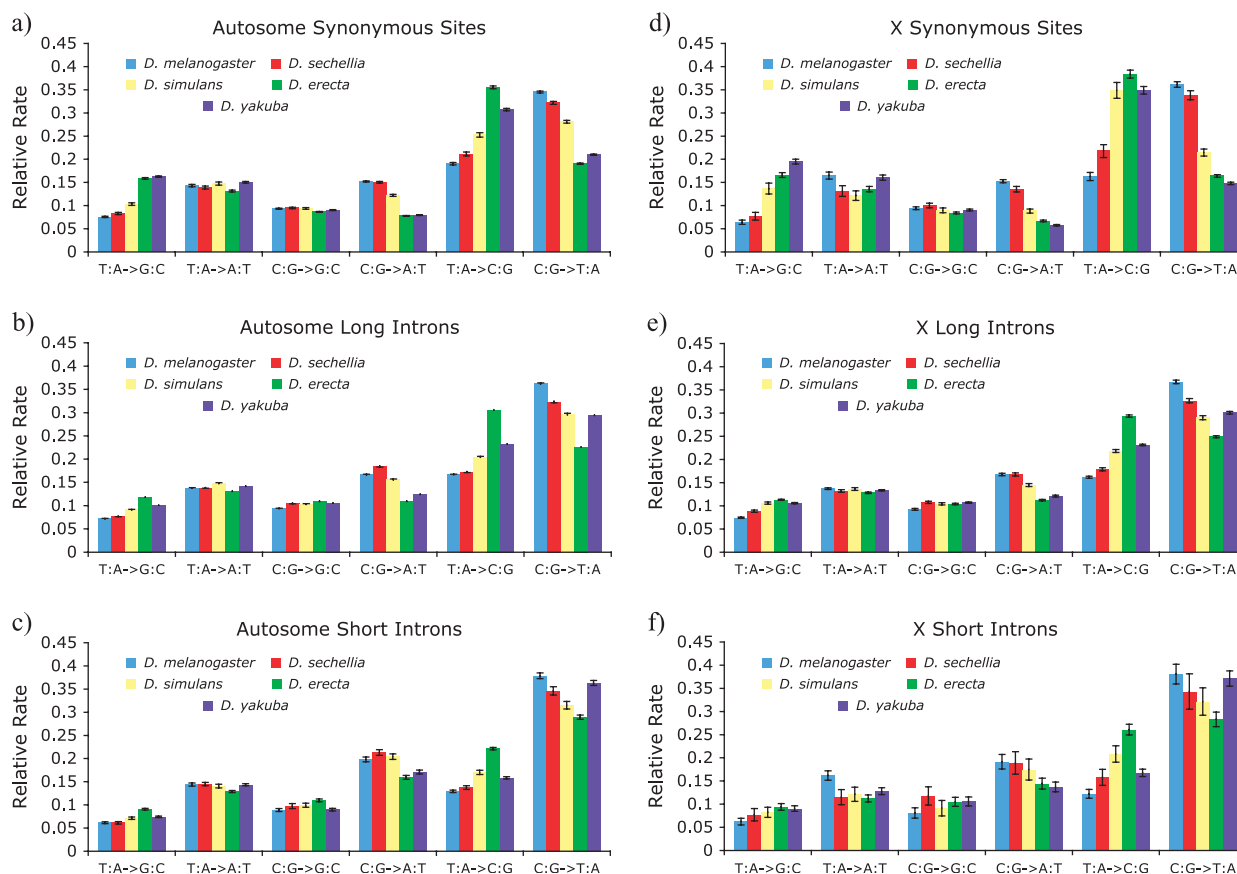


FIG. 1.—Rates of each of the six complementary pairs of single nucleotide substitutions in the five species of the *melanogaster* subgroup in (a) 4-fold degenerate synonymous sites on the autosomes, (b) long autosomal introns, (c) short autosomal introns, (d) 4-fold degenerate synonymous sites on the X chromosome, (e) long introns on the X chromosome, and (f) short introns on the X chromosome. Error bars denote standard error due to sampling.

we estimated lineage-specific transition matrices for X-linked and autosomal sequences separately, and within each of these chromosome sets, we estimated transition matrices for 4-fold degenerate synonymous sites, short introns, and long introns. These data are equivalent to those presented above, though they are grouped differently to highlight differences among sequence classes rather than among lineages.

There is striking variation in the relative rates of single nucleotide substitution among different sequence classes within a given species (fig. 1, supplementary fig. 6, Supplementary Material online). For both the X chromosome and the autosomes, the relative rates of transitions seem particularly variable, with GC-enriching transitions highest at 4-fold degenerate sites, at intermediate levels in long intron sequences, and lowest in short introns, encompassing 1.56- to 2.21-fold variation among sequence categories across the *melanogaster* subgroup. The increased rate of these transitions at 4-fold degenerate sites relative to long introns is statistically significant in all species on both the X and the autosomes ( $P \ll 0.001$ , 2-tailed *t*-test, all comparisons), with the exception of the *D. melanogaster* X chromosome. The increased rate of these transitions in long as compared with short introns is also statistically significant in all species on both the X and the autosomes ( $P < 0.005$ , 2-tailed *t*-test, all comparisons), with the exception of the *D. simulans* and *D. sechellia* X chromosomes.

In a parallel trend, the AT-enriching transitions show precisely the opposite pattern, although it is only statistically significant on the X and the autosomes in *D. yakuba* and *D. erecta* and on the autosomes in *D. melanogaster* ( $P \ll 0.001$ , 2-tailed *t*-test, all comparisons). In *D. simulans*, AT-enriching transitions are significantly increased at synonymous sites relative to long introns on the X and autosomes ( $P \ll 0.001$ , 2-tailed *t*-test, both comparisons), and in *D. sechellia* these transitions have a significantly higher rate in long versus short autosomal introns ( $P = 0.01$ , 2-tailed *t*-test). The rate of this transition varies 1.08- to 2.51-fold among sequence categories, depending on the species.

Interestingly, the rate of C:G  $\rightarrow$  A:T transversions appears to follow the same trend. This AT-enriching substitution is lowest at 4-fold degenerate sites, at intermediate levels in long introns, and is highest in short intron sequences and across all sites varies 1.23- to 2.61-fold depending on the species. The pairwise comparisons of the rates of this transversions between synonymous sites and long introns are statistically significant on both the X and autosomes in all species, and the long versus short intron comparison is significant on the autosomes in all species and on the X chromosome in *D. erecta* ( $P < 0.001$ , 2-tailed *t*-test, all comparisons). The reverse is true for the GC-enriching T:A  $\rightarrow$  G:C transversions, which shows significantly

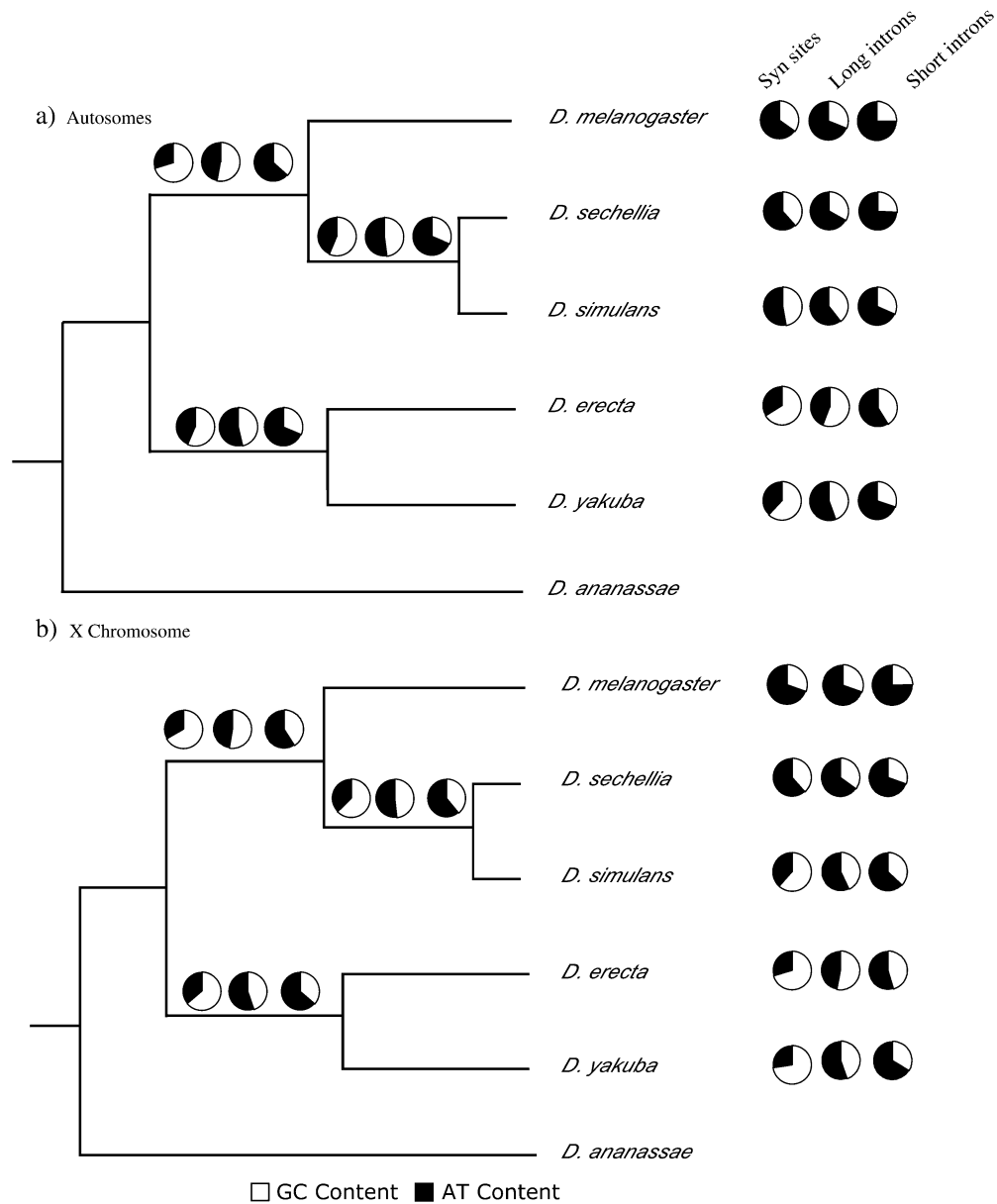


FIG. 2.—Expected stationary GC content at 4-fold degenerate synonymous sites (leftmost pie chart), long introns (middle pie chart), and short introns (rightmost pie chart) on the (a) autosomes and (b) X chromosome in extant species of *melanogaster* subgroup as well as at three internal nodes on the subgroup phylogeny. Note that substitutional rates and stationary GC content could not be estimated on the branch leading to *Drosophila ananassae* and the branch preceding split of the *melanogaster* species complex and the *Drosophila erecta*/*Drosophila yakuba* lineages because these branches are connected to the root.

higher rates in synonymous sites than in long introns in all species for both the X and the autosomes and significantly higher rates in long compared with short introns in the autosomes of all species and on the X chromosome of *D. erecta* and *D. yakuba* ( $P < 0.01$ , 2-tailed *t*-test, all comparisons). The relative rates of these transversions vary 1.30- to 2.98-fold among sequence categories.

These categorical differences in the relative rates of substitution, particularly those contributing greatly to changes in GC content, are perhaps best captured by stationary GC content, which is estimated using the stationary distribution of the transition matrix. Stationary GC content varies up to 2.4-fold among sequence types, depending on

the species. Consistent with the profile of single nucleotide substitutions, GC content is highest in coding sequences (ranging from 30% to 72% among species), at intermediate levels in long introns (31–56%), and is lowest in short introns (24–45%), for both X-linked and autosomal sequences (fig. 2). This difference in stationary GC content between long and short introns is consistent with the previously observed positive correlation between intron length and GC content (Haddrill et al. 2005).

In addition to comparing the relative rates of each of the six complementary pairs of single nucleotide substitutions, we can quantify variation in the total rate of single nucleotide substitution among sequence classes and

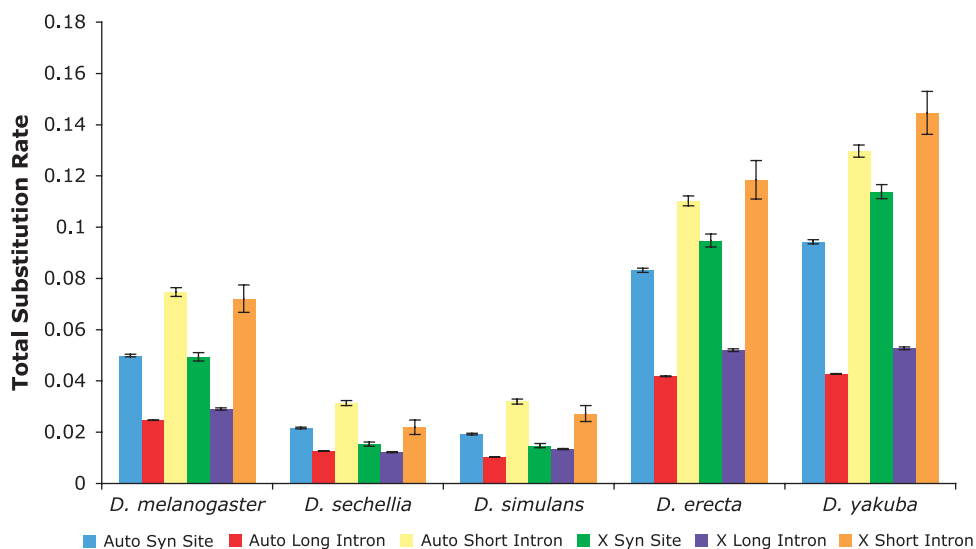


FIG. 3.—Total per-lineage substitution rate (per site) in synonymous and intronic sequences in the *melanogaster* subgroup. Error bars denote standard error due to sampling.

chromosomes. Among all species, for both the X chromosome and the autosomes, short introns appear to have the highest total substitution rate per lineage, followed by 4-fold degenerate synonymous sites and finally long intron sequences (fig. 3). The magnitude of the difference in total substitution rate among sequence classes appears consistent across species as well, with 4-fold degenerate synonymous sites evolving at roughly 70% of the rate of short introns on both the X and the autosomes, and with autosomal long introns evolving at 50% the rate of synonymous sites in autosomal sequences. However, the relative reduction in evolutionary rate of X-linked long introns relative to X-linked synonymous sites is highly variable among species, with reductions of ~40% to 55% in *D. melanogaster*, *D. erecta*, and *D. yakuba* but reductions of only ~10% to 20% in *D. simulans* and *D. sechellia* (fig. 3).

#### Intragenomic Variation in Rates and Patterns of Substitution: X versus Autosomes

Comparisons of the substitutional profiles of X-linked versus autosomal sequences are less straightforward. Within a given sequence class, there do not appear to be consistent differences in the relative rates of GC-enriching or AT-enriching substitutions between the X chromosome and the autosomes across species (supplementary fig. 6, Supplementary Material online). Differences in the total substitution rate between X-linked and autosomal sequences also show little consistency across species. Although in *D. erecta* and *D. yakuba*, X-linked sequences evolve at a 10–20% increased rate relative to autosomal sequences in each sequence category, the opposite is seen in *D. sechellia*, where X-linked sequences show a reduction in evolutionary rate of 5–30% (fig. 3). In *D. simulans* and *D. melanogaster*, 4-fold degenerate synonymous sites and short introns evolve less rapidly on the X chromosome (1–30% more slowly at synonymous sites, 4–15% more slowly at short introns), whereas long introns evolve

17–30% more quickly on the X than on the autosomes. Thus, although there is statistical support for estimating transition matrices for the X and autosomes separately based on our model fitting, the differences in substitutional patterns between the X and the autosomes are not suggestive of general trends, and will not be discussed further.

#### Nonequilibrium Base Composition

We compared the distribution of GC content observed in each of the terminal lineages in the *melanogaster* subgroup with the distribution of expected stationary GC content in each of our classes of sequence to identify departures from base composition equilibrium (fig. 4). Importantly, because our analysis was based on concatenated sequence data, the standard errors of the mean expected and observed GC content are due to sampling only. Although in *D. melanogaster* and *D. sechellia*, mean observed GC content is significantly higher than mean expected GC content in all sequence classes ( $P < 0.0001$ , all comparisons, 2-tailed  $t$ -test), in *D. simulans*, this is only the case for autosomal sequences and X-linked 4-fold degenerate synonymous sites (fig. 4). In contrast, base composition of short X-linked introns in *D. simulans* appear to be statistically indistinguishable from stationary frequencies, whereas long introns on the X chromosome in this species appear to have lower mean observed GC content than expected ( $P = 0.006$ , 2-tailed  $t$ -test). In *D. erecta* and *D. yakuba*, the pattern is slightly more complicated, with mean observed GC content at X-linked and autosomal 4-fold degenerate synonymous sites significantly higher than the mean expected GC content ( $P < 0.0001$ , both comparisons, 2-tailed  $t$ -test); this is also true of long X-linked introns in both of these species ( $P < 0.0001$ , both comparisons, 2-tailed  $t$ -test). However, in both species, mean observed GC content in long autosomal introns is significantly reduced relative to expectation ( $P < 0.0001$ , both comparisons, 2-tailed  $t$ -test; fig. 4). Short autosomal introns in *D. yakuba* have

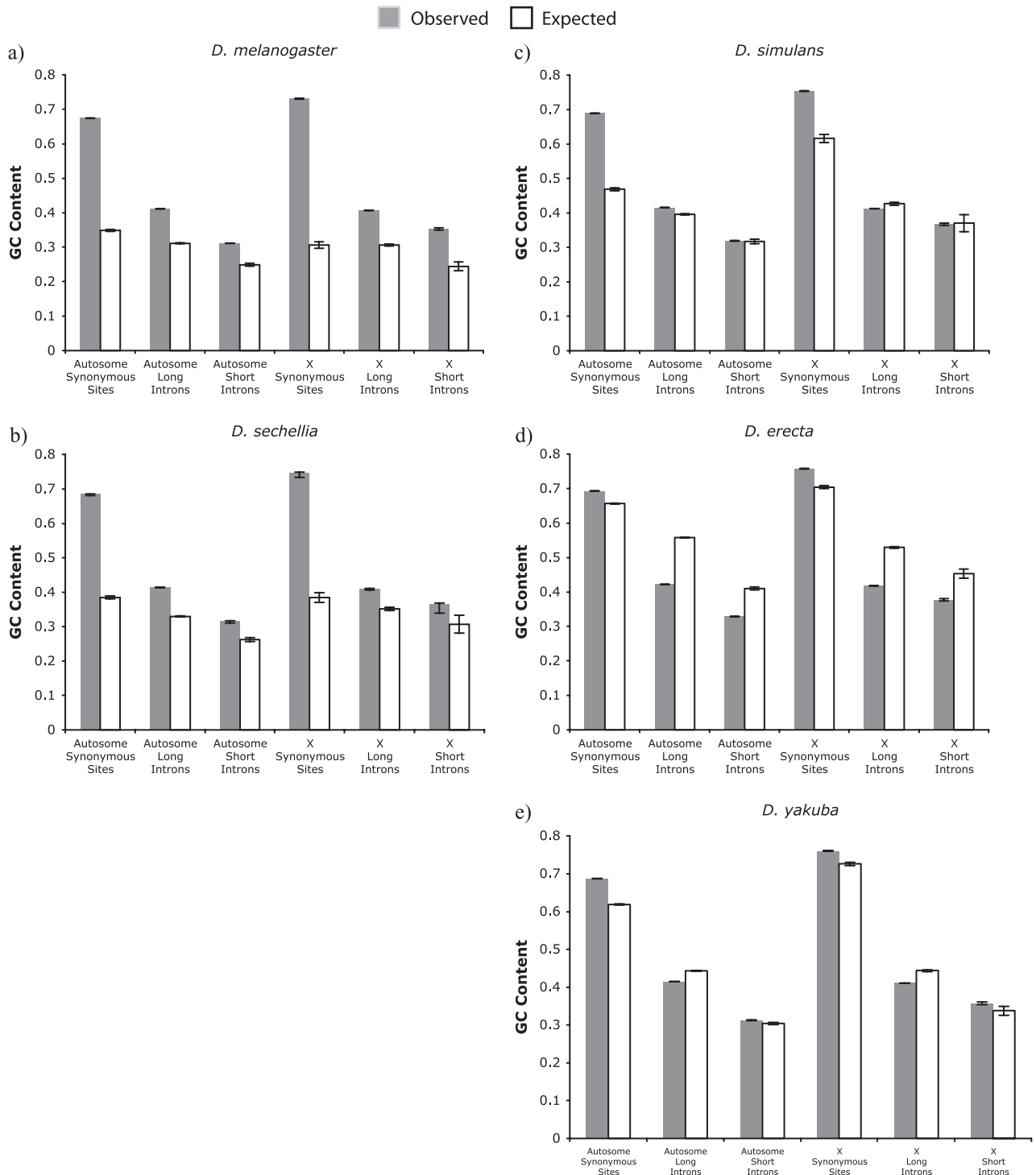


FIG. 4.—Comparison of observed GC content compared with expected stationary GC content given the stationary distributions of the estimated lineage-specific transition matrices in (a) *Drosophila melanogaster*, (b) *Drosophila sechellia*, (c) *Drosophila simulans*, (d) *Drosophila erecta*, and (e) *Drosophila yakuba*. Expected stationary GC content presented here is equivalent to those data presented in figure 2.

significantly higher observed GC content than expected ( $P = 0.02$ , 2-tailed  $t$ -test), though observed GC content in short autosomal introns in *D. erecta* is significantly lower than expected ( $P < 0.0001$ , 2-tailed  $t$ -test). Finally, although base frequencies of X-linked short introns in *D. yakuba* are not significantly different from equilibrium

base frequencies, in *D. erecta*, observed GC content in short X-linked introns is significantly higher than expected (fig. 4).

To investigate the evolution of base composition in these species in more depth, we used the estimated lineage-specific transition matrices to study GC-content dynamics as a function of evolutionary time (see Materials and

**Table 1**  
**GC Content Half-Life (Time to equilibrium<sup>a</sup>)**

	Autosome Synonymous Sites	Autosome Long Introns	Autosome Short Introns	X Synonymous Sites	X Long Introns	X Short Introns
<i>Drosophila melanogaster</i>	12.11 (60.90)	23.31 (77.51)	7.27 (19.21)	12.13 (65.58)	23.31 (77.51)	7.57 (26.06)
<i>Drosophila sechellia</i>	28.86 (141.79)	46.71 (144.42)	17.76 (43.18)	41.11 (213.25)	46.71 (144.42)	25.88 (65.51)
<i>Drosophila simulans</i>	35.55 (158.59)	62.11 (67.12)	18.33 (51.43)	53.52 (202.37)	62.11 (67.12)	21.90 (36.10)
<i>Drosophila erecta</i>	9.76 (17.87)	18.22 (68.58)	5.82 (17.61)	9.21 (22.36)	18.22 (68.58)	5.61 (16.47)
<i>Drosophila yakuba</i>	8.13 (22.50)	15.59 (23.62)	4.39 (NA <sup>b</sup> )	7.61 (13.47)	15.59 (23.62)	4.00 (3.81)

Because of uncertainty in speciation times for many lineages, time is given in units of terminal branch length. One can thus compare these metrics within species with relative ease. In addition, because the units of absolute time for *D. sechellia* and *D. simulans* are identical, as are the times for *D. erecta* and *D. yakuba*, GC content half-life and time to equilibrium can be compared within these sister species pairs.

<sup>a</sup> Equilibrium GC content defined as within 1% of the stationary GC content.

<sup>b</sup> NA corresponds to cases for which observed GC content is already within 1% of the stationary GC content.

Methods). Specifically, we estimated GC content half-life, or the amount of time required to reduce the difference between the current observed GC content and estimated stationary GC content by a factor of 2, for each species in the *melanogaster* subgroup (table 1). Similarly, we estimated time to equilibrium base composition, where we define equilibrium as being within one percent of the stationary GC content ( $f_{eq} = f_{stat} \pm 0.01$ ) (table 1). There is substantial variation in the rate of GC content decay as well as time to equilibrium among species and among sequence classes. Across species, consistent with the variation in rates of substitution among sequence classes, short introns show the shortest half-lives and times to base composition equilibrium, followed by 4-fold degenerate synonymous sites and finally long introns.

## Discussion

Rates of substitution depend on a number of population genetic parameters such as mutation rate, effective population size, and the coefficient of selection. As a consequence, heterogeneities in these parameters among and within species can lead to substantial variation in substitution rates. Here, we use a maximum likelihood framework to quantify the extent of heterogeneity in substitutional patterns in both inter and intragenomic contexts in the *melanogaster* group. It is important to note that this model does not take into account context-dependence of the substitution process, rate heterogeneity across sites, or incomplete lineage sorting and that all estimates of error are due to sampling only.

Additionally, there is ascertainment bias inherent in the coding and noncoding sequences used in this analysis. For the 4-fold degenerate synonymous sites, we restricted ourselves to those codons encoding conserved amino acids across the *melanogaster* group. In addition, the short and long intron sequences are, by necessity, those that remain alignable among these species as well. However, the effects of this ascertainment bias on inferences of substitutional rates and patterns are difficult to predict a priori. Moreover, all orthologous sequences within a given sequence category are subject to the same ascertainment bias, which suggests that our comparisons of substitutional patterns between genomes are still appropriate. Within genomes, it may very well be that different types of sequences are subject to different types of ascertainment bias. Synonymous sites might be comparatively easy to align

given that they are flanked by largely nondegenerate sites, whereas long introns might suffer from the greatest ascertainment bias. Indeed, our inferences of constraint may reflect this ascertainment bias to some extent. However, if ease of alignment were the sole underlying reason for heterogeneity in substitution rates, we would expect that synonymous sites should have the highest rate of substitution, which we do not find. Consequently, although ascertainment bias likely contributes to our observed heterogeneity in evolutionary rate among sequence classes, we do not believe that it is solely responsible for generating the patterns we observe.

## Lineage Specificity of Substitutional Patterns

Heterogeneity in rates of evolution have been reported previously, such as among more distantly related species with notable differences in extant base composition (Yang and Roberts 1995). In addition, analyses of subsets of genes or lineages in *Drosophila* certainly have suggested that rates and patterns of single nucleotide substitutions can vary among lineages (e.g., Eanes 1994; Akashi 1996; Takano 1998; Takano-Shimizu 2001; Akashi et al. 2006; Ko et al. 2006; Singh et al. 2006; Nielsen et al. 2007). We believe our results from *Drosophila* are particularly striking given the close evolutionary relationships among species in the *melanogaster* group and the consistency in base composition observed among extant species (supplementary fig. 5, Supplementary Material online). In addition, we take advantage of a genome-scale data set, which suggests that the observed lineage specificity of substitutional patterns is not driven by a subset of loci in the genome.

Our analysis suggests that factors influencing base composition evolution across the *melanogaster* group phylogeny are highly dynamic. In particular, it appears as if patterns of single nucleotide substitution are comparatively more AT-biased in the *melanogaster* species complex, particularly in *D. melanogaster* and *D. sechellia*, than they are in the *yakuba/erecta* clade (figs. 1 and 2). Moreover, this reduction in stationary GC content appears to be a derived feature of the *melanogaster* species complex, as inferred values of stationary GC content at the internal nodes of the *melanogaster* group phylogeny are comparatively higher (fig. 2). Because we are investigating the substitutional process, it is difficult to disentangle

the relative contributions of mutation, random genetic drift, BGC and natural selection. However, one possible explanation for the observed patterns of base composition in the *melanogaster* group is a shift in mutational patterns in the ancestor of the *melanogaster* species complex toward increased AT.

The plausibility of the mutational shift hypothesis is comparatively difficult to evaluate in the absence of large-scale polymorphism data from several species in the *melanogaster* group. However, there have been some suggestions that mutational patterns are relatively labile in *Drosophila*. For instance, species in the *D. willistoni* and *D. saltans* clades have a dramatically reduced GC content (Anderson et al. 1993; Rodriguez-Trelles et al. 1999, 2000a, 2000b; Powell et al. 2003; Heger and Ponting 2007; Vicario et al. 2007), particularly in coding sequences, which appears to be partially explained by a shift in background substitutional patterns in this clade (Singh et al. 2006). In addition, other *Drosophila* species show patterns of polymorphism and divergence that are consistent with shifts in mutational patterns (Takano-Shimizu 1999, 2001; Kern and Begun 2005; Akashi et al. 2006), and strikingly, mutation profiles have recently been shown to be significantly different between the recently diverged *D. melanogaster* and *D. sechellia* (Singh et al. 2007). It thus seems possible that the observed reduction in stationary GC content in the *melanogaster* species complex based on substitutional patterns results from a lineage-specific shift in mutational patterns in this clade. Importantly, this model is difficult to distinguish from a model in which there has been a shift in mutation–selection balance in the *melanogaster* species complex. Because mutation is AT-biased, a simple relaxation of constraint in these lineages or a reduction in the efficacy of natural selection could in principle generate this pattern as well.

Alternatively, these data may also be consistent with a prominent role of BGC in genome evolution. Gene conversion may be GC-biased in *Drosophila* (Galtier et al. 2006), and there is growing support for a role of BGC in *Drosophila* genome evolution (Bartolome et al. 2005; Galtier et al. 2006; Haddrill and Charlesworth 2008). Differences in effective population size among lineages could result in heterogeneity in base composition among species. Under this model, species with the smallest effective size would show the weakest bias in substitutional patterns toward GC. However, there is little consistent empirical data on effective population size among *Drosophilids*, even among well-characterized species such as *D. melanogaster* and *D. simulans*. Some of the previous inferences of effective population size based on polymorphism data are suggestive of a reduced effective population size in *D. melanogaster* relative to *D. simulans* (e.g., Morton et al. 2004), whereas other such studies indicate comparable effective sizes (e.g., Nolte and Schlotterer 2008). The lack of available empirical data on historical population sizes of the species studied herein makes it challenging to evaluate the plausibility of this model, and it does remain a formal possibility. Genome-scale polymorphism data from representative lineages in the *melanogaster* group will likely provide key insight into the applicability of this model in the future.

## Sequence-Specificity of Substitutional Patterns

Rates of single nucleotide substitution are known to vary intragenomically in *Drosophila*. In particular, different classes of sequence in *Drosophila* appear to be subject to varying degrees of selective constraint, which results in variable rates of evolution among different types of genomic sequence. There is a growing body of literature suggesting that noncoding sequences are constrained in *Drosophila* (Bergman and Kreitman 2001; Halligan et al. 2004; Marais et al. 2005; Halligan and Keightley 2006). Moreover, introns of different lengths are differentially constrained (Haddrill et al. 2005).

To expand upon previous analyses, we systematically characterized the extent and magnitude of variation in rates and patterns of single nucleotide substitution at a genomic scale in the five terminal lineages of the *melanogaster* subgroup. Our analysis clearly reveals significant differences in substitutional rates and patterns among sequence classes. It is important to note that our sequence classes do differ with respect to their GC content (supplementary fig. 5, Supplementary Material online). Although there is some evidence that rates of substitution are related to GC content in *Drosophila* (Haddrill et al. 2005), we do not believe that the context-dependence of substitutional patterns in *Drosophila* is likely to be as strong as is found in humans (Arndt et al. 2005), for instance, given the lack of highly structured GC content in the *Drosophila* genome. It is certainly possible that some of the patterns we find within our sequence categories are driven at least in part by differences in GC content, and future work will be required to assess the relative contributions of sequence class and GC content to patterns of single nucleotide substitution in *Drosophila*.

Our results indicate that 4-fold degenerate sites show strongly GC-biased substitutional patterns on both the X chromosome and the autosomes, particularly in non-*melanogaster* species. This is consistent with selection on codon bias in the *melanogaster* group on several levels. First, it is generally accepted that with the exception of at most one codon, preferred codons in *D. melanogaster* are G- or C-ending (Akashi 1995; Duret and Mouchiroud 1999). As a consequence, increased rates of GC-enriching substitutions and decreased rates of AT-enriching substitutions at synonymous sites relative to intronic sequences (figs. 1 and 2) are entirely consistent with selection pressure favoring biased codon usage. Second, it is well documented that the *D. melanogaster* lineage has experienced a genomic reduction in codon bias (Akashi 1995, 1996; McVean and Vieira 2001; Bauer DuMont et al. 2004; Akashi et al. 2006; Nielsen et al. 2007; Singh et al. 2007), and it is therefore unsurprising that substitutional patterns at synonymous sites in *D. melanogaster* deviate from those observed across the remainder of the phylogeny. Finally, that stationary GC content at 4-fold degenerate sites is inferred to be greater on the X chromosome than on the autosomes in *D. erecta*, *D. yakuba*, and *D. simulans* is further consistent with selection on codon bias, given previous observations that codon bias of X-linked genes is consistently elevated in *Drosophila* (Comeron et al. 1999; Hambuch and Parsch 2005; Singh et al. 2005, 2008).

Beyond examining patterns of evolution in various classes of genomic sequence, these lineage-specific transition matrices can be used to compare overall substitution

rates between coding and intronic sequences, and to make inferences about selective constraint in *Drosophila*. Strikingly, long introns show the lowest rate of single nucleotide substitution, followed by 4-fold degenerate synonymous sites, and short introns show the most rapid rate of evolution across all species for both the X chromosome and the autosomes. This is largely consistent with previous reports, which have suggested that introns are more highly constrained than synonymous sites (Bergman and Kreitman 2001; Andolfatto 2005; Haddrill et al. 2005; Marais et al. 2005; Halligan and Keightley 2006). However, our observation that short intron substitution rates significantly exceed those at 4-fold degenerate synonymous sites (fig. 3) differs slightly from previous work, which suggested that synonymous sites and short introns evolve at statistically indistinguishable rates (Haddrill et al. 2005). Given that our estimates of pairwise divergence at intronic sites are quite close to those reported previously (Haddrill et al. 2005), the difference between these two studies is driven by divergence at synonymous sites. Importantly all synonymous sites were used in the previous study, whereas the current analysis was limited to 4-fold degenerate synonymous sites. This sampling difference may underlie the reduced rate of synonymous sites divergence that we observe.

Interestingly, the following considerations suggest that rates of evolution at short introns are virtually identical to neutral expectation. Given a divergence time of 2.3 million years (My) between *D. melanogaster* and its sister species, and approximate divergence times between *D. simulans* and *D. sechellia* of 800,000 years (Russo et al. 1995), we can estimate the rate of substitution per site per million years. At short introns, rates of substitution in these three species range from 0.027 to 0.40 substitutions/site/My on the X chromosome and the autosomes, whereas previous estimates of the neutral substitution rate in *Drosophila* average 0.033 substitutions/site/My (Singh and Petrov 2004; supplementary fig. 7, Supplementary Material online). Although this observation is tantalizing, it is important to note that the previous estimates were based on a small number of unconstrained sequences, and that the methodology of estimating divergence also differs between the current and previous study. Moreover, a recent analysis of pairwise divergence between *D. melanogaster* and *D. simulans* in transposable elements suggests a slightly higher neutral substitution rate of 0.40 substitutions/site/My (Wang et al. 2007). Our results thus suggest that at least with respect to rates of single nucleotide substitution, short introns evolve in a manner that is most consistent with neutrality relative to other sequence classes, whereas synonymous sites are consistently more constrained and long introns are more constrained yet.

#### Substitutional Patterns in *D. melanogaster*

One of the more striking results from this analysis is the similarity in the relative rates of single nucleotide substitutions in *D. melanogaster* across different classes of genomic sequence (supplementary fig. 6, Supplementary Material online). This contrasts with the patterns observed in the remaining species in the *melanogaster* subgroup (with the exception of *D. sechellia*), which show greater variation in relative rates

of GC-enriching and AT-enriching substitutions across sequence classes. These classes of substitution vary only 1.1- to 1.6-fold in *D. melanogaster* (depending on the single nucleotide substitution), but vary 1.5- to 3.0-fold in *D. simulans*, *D. yakuba*, and *D. erecta*. One potential explanation for this observation is that the *D. melanogaster* lineage has experienced a reduction in effective population size, which reduces the efficacy of natural selection at a genomic scale. This has often been suggested as contributing to the genomic reduction in codon bias in this lineage (Akashi 1995, 1996; McVean and Vieira 2001; Bauer DuMont et al. 2004; Akashi et al. 2006; Nielsen et al. 2007; Singh et al. 2007). Our substitution rate data thus tentatively support a model of reduced efficacy of natural selection specifically on the *D. melanogaster* lineage, although under such a model one might expect generally increased of substitution on this lineage, which we do not observe (supplementary fig. 7, Supplementary Material online). Further investigation of this issue is thus warranted.

#### Toward a Model of Base Composition Evolution

Comparing inferred stationary GC content with observed GC content in extant sequences revealed marked departures from equilibrium in many cases (fig. 5), consistent with previous reports (Singh et al. 2004, 2007; Kern and Begun 2005; Akashi et al. 2006). Although synonymous sites show the largest deviation from equilibrium in all species, there is marked variation among species in the extent of departure from equilibrium base composition. Most notably, *D. melanogaster* and *D. sechellia* show the most dramatic and significant differences between expected and observed GC content, with differences between observed and expected GC content ranging from 5% to 42%. *D. simulans* shows a similar pattern, albeit to a lesser extent, as deviations from expectation only range from 0.1% to 22%. These observations may reflect the effects of the shift in mutational patterns derived in the *melanogaster* species complex toward AT-enriching mutations, or a shift in mutation–selection balance favoring the inherently AT-biased mutation process in this clade. The differences observed among these three species may reflect differences in effective population size, as there is some evidence that the effective sizes of *D. melanogaster* and *D. sechellia* appear smaller than that of *D. simulans* (Morton et al. 2004, though see also Nolte and Schlotterer 2008). In contrast, observed GC content values in *D. yakuba* and *D. erecta* are comparatively closer to equilibrium, deviating from expectation by only 2–14%, though this is not entirely unexpected given that substitutional patterns in these lineages are more similar to the inferred ancestral substitutional patterns.

Our time to equilibrium calculation suggest that base composition evolution is comparatively slow. With the exception of short autosomal introns in *D. yakuba*, where extant base composition is by definition already at equilibrium, the smallest time to reach base composition equilibrium, which is seen in short X-linked introns in *D. yakuba*, is slightly less than four time units (table 1), which indicates that it would take almost four times the speciation time for *D. yakuba*. Given that the divergence time between *D. yakuba* and *D. erecta* is on the order of 5 My, this suggests that it would

require nearly 20 My for base composition of the sequences to reach equilibrium with respect to patterns of single nucleotide substitution. This is at least three times the evolutionary time separating *D. yakuba* from *D. melanogaster*, representing the most distant relations in the *melanogaster* subgroup. As a consequence, it appears as though patterns of single nucleotide substitution are shifting across lineages on a timescale that is more rapid than base composition evolution can accommodate, such that base composition is unlikely to reach equilibrium with respect to substitutional patterns in *Drosophila*.

### Conclusions and Future Directions

Our analysis provides rigorous statistical evidence that *Drosophila* species differ significantly in their rates and patterns of single nucleotide substitution at a genomic scale, and that even within genomes, substitutional patterns vary significantly by chromosome and sequence type. Relative rates of each of the six complementary pairs of single nucleotide substitution vary up to 3-fold, and equilibrium GC content varies up to 2.4-fold, among species within a given sequence category. In addition, within species, relative rates of single nucleotide substitution vary up to 3-fold across sequence types, and equilibrium GC content varies approximately 2.4-fold. Thus, the degree of heterogeneity in substitutional patterns within genomes is similar in magnitude to the degree of heterogeneity among genomes in the *D. melanogaster* species subgroup.

These data highlight the variability of the substitutional process across the genome, echoing previous reports. Our results are consistent with the previous inference that some introns are selectively constrained and further confirm that long intron sequences are more highly constrained than synonymous sites. Moreover, our results hint at the possibility that short introns are evolving in manner that is more consistent with neutrality than evolutionary rates at synonymous sites. These data thus increase our understanding of the distribution of selective constraint across the genome in *Drosophila* and will hopefully serve to help better inform our choices of neutral models of sequence evolution.

Most notably, we find strong statistical evidence for the lineage specificity of the substitution process. Specifically, individual lineages show significant differences in both rates and patterns of single nucleotide substitution, which have the potential to have profound effects on the evolution of base composition. Overall, our results support a shift in substitutional patterns in the *melanogaster* species complex coupled with differences in effective population size among species within this complex. This may ultimately be due to a mutational shift toward increased AT content, or a shift in mutation–selection balance due to relaxed constraint or a reduced efficacy of natural selection, in this complex. As more polymorphism data for *Drosophila* become available, it will become possible to explicitly test these hypotheses.

Finally, we find substantial support for nonequilibrium base composition, though the direction and magnitude of the deviation from equilibrium vary among comparisons. These results have marked implications for evolutionary analyses. Notably, phylogenetic analyses of orthologous DNA sequen-

ces typically assume constancy of the substitution matrix across the entire phylogeny and also assume that base composition of these sequences is at equilibrium. Our results thus indicate that both of these assumptions are violated in both coding and intronic sequences for species in the *D. melanogaster* subgroup. The effects of these violations on evolutionary inference unfortunately remain unclear, and this will likely be a fruitful line of future investigation.

### Supplementary Material

Supplementary figures 1–7 and supplementary table 1 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgments

The authors thank R. Durrett for his assistance on the time to equilibrium calculations and V. Bauer Dumont for providing the intronic sequences. The authors are also grateful to L. Duret, M. Hamblin, V. Bauer Dumont, and N. Clark for insightful feedback on this work. Comments from our handling editor, Jeff Thorne, as well as two anonymous reviewers substantially improved the quality of this manuscript. This work was supported in part by an National Institutes of Health National Research Service Award (grant number 1F32GM080944-01 to N.D.S., C.F.A., and A.G.C.).

### 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, 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.
- 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. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*. 437:1149–1152.
- Arndt PF, Hwa T. 2005. Identification and measurement of neighbor dependent nucleotide substitution processes. *Bioinformatics*. 21:2322–2328.
- Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density and telomere-specific effects. *J Mol Evol*. 60:758–763.
- Avery PJ. 1984. The population genetics of haplo-diploids and X-linked genes. *Genet Res*. 44:321–341.
- Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics*. 174:2045–2059.
- Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. *Mol Biol Evol*. 25:1639–1650.
- Bartolome C, Maside X, Yi S, Grant AL, Charlesworth B. 2005. Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*. *Genetics*. 169:1495–1507.

- 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.
- Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors) 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:2534–2559.
- Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics*. 156:1879–1888.
- Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res*. 11:1335–1345.
- Betancourt AJ, Kim Y, Orr HA. 2004. A pseudohitchhiking model of x vs. autosomal diversity. *Genetics*. 168:2261–2269.
- Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in *Drosophila*. *Proc Natl Acad Sci USA*. 99:13616–13620.
- Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A test for faster X evolution in *Drosophila*. *Mol Biol Evol*. 19:1816–1819.
- Bhutkar AV, Russo S, Smith TF, Gelbart WM. 2007. Genome-scale analysis of positionally relocated genes. *Genome Res*. 17:1880–1889.
- Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat*. 130:113–146.
- 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.
- Counterman BA, Ortiz-Barrientos C, Noor MAF. 2004. Using comparative genomic data to test for fast-X evolution. *Evolution*. 58:656–660.
- Davis JC, Brandman O, Petrov DA. 2005. Protein evolution in the context of *Drosophila* development. *J Mol Evol*. 60:774–785.
- Drosophila* 12 Genomes Consortium (244 co-authors). 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*. 450:203–218.
- Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.
- 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.
- Eanes WF. 1994. Patterns of polymorphism and between species divergence in the enzymes of central metabolism. Pp. 18–28. in B Golding, editor. *Non-neutral evolution: theories and molecular data*. Chapman and Hall, New York.
- Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*. 8:689–698.
- 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.
- Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics*. 172:221–228.
- Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol Biol Evol*. 25:1825–1834.
- Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biol Lett*. 4:438–441.
- Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent on length and GC content. *Genome Res*. 6:R:67.
- Haerty W, Jagadeeshan WS, Kulathinal RJ, et al. (11 co-authors) 2007. Evolution in the fast lane: rapidly evolving sex-and reproduction-related genes in *Drosophila* species. *Genetics*. 177:1321–1335.
- Halligan DL, Eyre Walker AC, Andolfatto P, Keightley PD. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res*. 14:273–279.
- Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res*. 16:875–884.
- Hambuch TM, Parsch J. 2005. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. *Genetics*. 170:1691–1700.
- Heger A, Ponting C. 2007. Variable strength of translational selection among twelve *Drosophila* species. *Genetics*. 177:1337–1348.
- Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics*. 160:595–608.
- 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.
- Kliman RM, Hey J. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol*. 10:1239–1258.
- Ko WY, Piao SF, Akashi H. 2006. Strong regional heterogeneity in base composition evolution on the *Drosophila* X chromosome. *Genetics*. 174:349–362.
- Kohn MH, Fang S, Wu CI. 2004. Inference of positive and negative selection on the 5' regulatory regions of *Drosophila* genes. *Mol Biol Evol*. 21:374–383.
- Larracuent AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends Genet*. 24:114–123.
- Lazzaro BP. 2008. Natural selection on the *Drosophila* antimicrobial immune system. *Curr Opin Microbiol*. 11:284–289.
- Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22:1345–1354.
- Liu G, Li H. 2008. The correlation between recombination rate and dinucleotide frequency in *Drosophila melanogaster*. *J Mol Evol*. 67:358–367.
- Lobry JR, Lobry C. 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol Biol Evol*. 16:719–723.
- Lyko F, Ramashoye BH, Jaenisch R. 2000. DNA methylation in *Drosophila melanogaster*. *Nature (Lond)*. 408:538–540.
- Marais G, Domazet-Loso T, Tautz D, Charlesworth B. 2004. Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. *J Mol Evol*. 59:771–779.
- 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.

- Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in *Drosophila*. *Genetics*. 170:481–485.
- McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc Natl Acad Sci USA*. 104:4996–5001.
- McBride CS, Arguello JR. 2007. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*. 177:1395–1416.
- McVean GAT, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics*. 157:245–257.
- Moriyama EN, Powell JR. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol*. 13:261–277.
- 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.
- 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.
- Nolte V, Schlotterer C. 2008. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. *Genetics*. 178:405–412.
- Ometto L, De Lorenzo D, Stephan W. 2006. Contrasting patterns of sequence divergence and base composition between *Drosophila* introns and intergenic regions. *Biol Lett*. doi:10.1098/rsbl.2006.0521.
- Petrov DA, Hartl DL. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci USA*. 96:1475–1479.
- Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet*. 2:1634–1647.
- 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.
- Proschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics*. 174:893–900.
- 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, Tarrío R, Ayala FJ. 1999. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics*. 153:339–350.
- Rodriguez-Trelles F, Tarrío R, Ayala FJ. 2000b. Evidence for a high ancestral GC content in *Drosophila*. *Mol Biol Evol*. 17:1710–1717.
- Rodriguez-Trelles F, Tarrío R, Ayala FJ. 2000a. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J Mol Evol*. 50:1–10.
- Russo CAM, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of Drosophilid species. *Mol Biol Evol*. 12:391–404.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nat Genet*. 39:1461–1468.
- Schlenke TA, Begun DJ. 2003. Natural selection drives *Drosophila* immune system evolution. *Genetics*. 164:1471–1480.
- 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, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. 2007. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol Biol Evol*. 24:2687–2697.
- Singh ND, Davis JC, Petrov DA. 2005. X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. *Genetics*. 171:145–155.
- Singh ND, Larracuente AM, Clark AG. 2008. Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Mol Biol Evol*. 25:454–467.
- Singh ND, Petrov DA. 2004. Rapid sequence turnover at an intergenic locus in *Drosophila*. *Mol Biol Evol*. 21:670–680.
- Stark A, Lin MF, Kheradpour P, et al. (43 co-authors) 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*. 450:219–232.
- Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc Natl Acad Sci USA*. 98:7375–7379.
- Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet*. 3:137–140.
- Takano TS. 1998. Rate variation of DNA sequence evolution in the *Drosophila* lineages. *Genetics*. 149:959–970.
- 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.
- Thornton K, Bachtrog D, Andolfatto P. 2006. X chromosomes and autosomes evolve at similar rates in *Drosophila*: no evidence for faster-X protein evolution. *Genome Res*. 16:498–504.
- Thornton K, Long M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol Biol Evol*. 19:918–925.
- Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol Biol*. 7:226.
- Vinogradov AE. 2001. Intron length and codon usage. *J Mol Evol*. 52:2–5.
- Wang J, Keightley PD, Halligan DL. 2007. Effect of divergence time and recombination rate on molecular evolution of *Drosophila* INE-1 transposable elements and other candidates for neutrally evolving sites. *J Mol Evol*. 65:627–639.
- Yang ZH, Roberts D. 1995. On the use of nucleic-acid sequences to infer early branchings in the tree of life. *Mol Biol Evol*. 12:451–458.
- Zhang Z, Hambuch TM, Parsch J. 2004. Molecular evolution of sex-biased genes in *Drosophila*. *Mol Biol Evol*. 21:2130–2139.
- Zhang Z, Parsch J. 2005. Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression. *Mol Biol Evol*. 22:1945–1947.
- Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B. 2007. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature*. 450:233–238.
- Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Evol*. 66:395–404.

Jeffrey Thorne, Associate Editor

Accepted April 1, 2009