Patterns of Polymorphism and Divergence from Noncoding Sequences of *Drosophila melanogaster* and *D. simulans*: Evidence for Nonequilibrium Processes

Andrew D. Kern and David J. Begun

Center for Population Biology, University of California, Davis

Despite the fact that *D. melanogaster* and *D. simulans* have been the central model system for molecular population genetics, few data are available for noncoding regions. Here, we present an analysis of population genetic data from intergenic regions and comparisons of these data to previously collected data from introns and exons. Polymorphisms and fixations were categorized as A/T to G/C or G/C to A/T changes and were polarized by inferring the ancestral state using both parsimony and maximum likelihood. Noncoding fixations in both *D. melanogaster* and *D. simulans* were consistent with equilibrium base-composition evolution. However, polarized noncoding polymorphisms, revealed a different pattern. Although A/T to G/C and G/C to A/T polymorphisms in *D. simulans* were consistent with equilibrium, we observed a highly significant dearth of A/T to G/C polymorphisms in *D. melanogaster* introns but not in intergenic sequences. Such data could be explained by recent evolution of mutational biases associated with transcription or by lineage-specific selection on base composition. These data reveal the complexity of evolutionary processes acting even on noncoding DNA in *Drosophila*.

Introduction

Comparison of polymorphism and divergence in closely related species of *Drosophila* has been an effective strategy for testing population genetic hypotheses on the causes of variation (Hudson, Kreitman, and Aguadé 1987; Berry, Ajokia, and Kreitman 1991; Begun and Aquadro 1992; Langley et al. 1993; Akashi 1994). However, most such data come from protein-coding regions. Basic evolutionary questions about noncoding regions, including the relative importance of mutation and selection to the evolution of such regions in *Drosophila*, are poorly understood. Because noncoding DNA represents the great majority of most eukaryotic genomes, a population genetics–based understanding of such DNA would be a significant advance for evolutionary genomics. The investigation of noncoding regions also has potential practical implications. A current focus of genomics is to identify functional elements in the noncoding portions of genomes. Generally this goal relies on identifying conserved regions among distantly related genomes (e.g., Bergman and Kreitman 2001; Webb et al. 2002). The idea that such regions are functional follows from the neutral theory of molecular evolution (Kimura 1983), which posits that functional constraint and evolutionary rate are negatively correlated. However, such analyses disregard the potential contributions of mutational heterogeneity across genomic regions and lineages.

In this paper, we describe patterns of polymorphism and divergence in two types of noncoding DNA, nontranscribed intergenic regions and introns, from *Drosophila melanogaster* and *Drosophila simulans*. Additionally, we compare these data to synonymous site data from exons. Our goals are to investigate whether transcription is associated with different patterns of mutation or selection and to compare patterns of variation in nontranscribed DNA from two lineages. Two observations motivate this goal. First, Akashi (1996) used divergence data to infer that intron base composition is at equilibrium in *D. melanogaster* and *D. simulans*. Second, Takano-Shimizu (2001) suggested that noncoding sequences along the lineage between *D. yakuba* and *D. melanogaster* had undergone a shift in mutational bias towards A/T nucleotides. One interpretation of these results is that the mutation process differs between intron and intergenic sequences in *Drosophila*. A possible explanation for such differences is that introns are transcribed, whereas most intergenic regions, presumably, are not. Indeed, transcription appears to have a significant impact on patterns of base composition in mammals, most likely as a consequence of germline transcription-coupled repair (Green et al. 2003; Majewski 2003). Although there is no evidence of transcription-coupled repair in flies (de Cock et al. 1992; van der Helm et al. 1997; Sekelsky, Brodsky, and Burtis 2000), other properties/effects of transcription could affect mutation biases or rates.

Materials and Methods

Fly Stocks/Sequencing

New *D. simulans* sequences reported here were collected from inbred lines derived from flies collected in the Wolfskill Orchard in Winters, California (Begun and Whitley 2000). New *D. melanogaster* intergenic sequences are from isofemale lines collected from Malawi and provided to us by W. O. Ballard. *D. yakuba* sequence is from one isofemale line, S180, from the Umea Stock Center. In all cases, direct sequencing of PCR products in both directions were determined on an ABI 3700 sequencer. Raw data were analyzed using phred (Ewing et al. 1998) and phrap (Ewing and Green 1998). Sequences were then examined by eye using Consed (Gordon, Abajian, and Green 1998). Polymorphic sites in multiple alignments were verified using the MACE package of scripts (http://ludwig.ucdavis.edu/MACE/). The intergenic loci sampled were designed to be a random sample of nontranscribed, noncoding DNA and were taken from...
the GadFly version 3.1 annotation of the *D. melanogaster* genome. In total, we sequenced seven intergenic loci, which we have designated according to cytological location. Inter1C1FR1 corresponds to the scaffold AE003490 from positions 162425 to 162954. Inter1C1FR6 is from AE003490 at 165475 to 165587. Both of these regions are approximately 3 kb from the nearest annotated gene. Inter21F corresponds to positions 286303 to 287103 on the scaffold AE003587 (approximately 3 kb from the nearest annotated gene). Inter49F comes from AE003819 positions 187258 to 187939 (approximately 1 kb from the nearest annotated gene). Inter65A corresponds to AE003563 from positions 103969 to 104613 (approximately 1 kb from the nearest annotated gene). Inter84E is from AE003676 positions 101165 to 102088 (approximately 500 bp from the nearest annotated gene). Inter86C is from AE003688 positions 215272 to 215768 (approximately 200 bp from the nearest annotated gene). Average local recombination rates for sampled intergenic loci were estimated by fitting a fourth order polynomial to the comparison of genetic and physical maps (Kliman and Hey 1993; data not shown). Recombination rates for intergenic sequences (2.43 cM/mb) were roughly equivalent to average recombination rates for the genes from which we have intronic/synonymous site data (2.11 cM/mb). New sequences reported in this paper can be found in GenBank under accession numbers AY757957 to AY758069. Alignments are available from A.D.K. upon request.

Intron Sequences

To obtain a lineage-specific picture of polymorphism and divergence from intronic regions, loci for which polymorphism data existed from *D. melanogaster* and *D. simulans* and which also had at least one outgroup sequence from *D. yakuba* were assembled from GenBank. In total, 49 introns from 24 loci (~5.5 kb) throughout the genome satisfied these criteria (tables 1 and 2). References for these sequences can be found in the publications listed in table 1. Most sampled introns were small, which is consistent with the heavily tailed distribution of intron sizes found in the *D. melanogaster genome* (Adams et al. 2000; Parsch 2003).

Sequence Analysis

All alignments were done using either ClustalW version 1.82 (Thompson et al. 1994) and/or DIALIGN version 2.2 (Morgenstern 1999) and were subsequently edited by hand. As introns are often difficult to align, we anchored each multiple alignment in flanking coding sequences. This provided much better sequence alignment for indel-rich introns. Intergenic sequences were easily aligned. Sites that showed indel polymorphism within or divergence between species were excluded from further analysis.

Many polymorphism and divergence statistics were estimated using software developed by A.D.K. Source code in Smalltalk is available upon request. To test the null hypothesis of neutral evolution, we performed multilocus Hudson-Kreitman-Aguadé tests (Hudson, Kreitman, Aguadé 1987) on intergenic and intronic data separately. These tests were carried out using a version of Jody Hey’s HKA program (Kliman et al. 2000) that we compiled to handle a larger number of loci. Critical values reported here for HKA tests were based on 10,000 neutral coalescent simulations rather than the standard chi-square distribution.

Parsimony and maximum-likelihood analyses were performed on a site-by-site basis to infer ancestral states for the common ancestor of *D. melanogaster* and *D. simulans*. Inferred ancestral states were used to identify lineage-specific fixations and to infer ancestral states for polymorphisms. For likelihood reconstruction, we used PAML (Yang 1997) under a number of substitution models. For these reconstructions, we assumed that within-species genealogies are star shaped. This could be a problem if one were indirectly inferring the number of polymorphisms within species because the assumption of a star-shaped genealogy would inflate the number of mutations inferred on terminal branches. However, because we are interested only in the internal node of the genealogy, which represents the most recent common ancestor of *D. melanogaster* and *D. simulans*, the star-phylogeny assumption is not expected to have a large effect on our inferences. Indeed, any error caused by this assumption appears to have had little effect on our conclusions, as we obtained qualitatively similar results independent of the mode of ancestral reconstruction employed. Furthermore, only minor differences were observed for different likelihood models (data not shown), most plausibly as a function of the short divergence times. Results reported here are from maximum-likelihood reconstructions under the general time reversible (GTR) model (Tavaré 1986), where both the transition/transversion rate and shape parameter for rate heterogeneity were estimated from the data. For parsimony reconstructions, only sites for which an unambiguous ancestral state could be inferred were considered. Although parsimony may lead to inaccurate ancestral inference over large evolutionary distances, it should provide a reliable criterion over the short divergence times between taxa considered here (Yang, Kumar, and Nei 1995). All parsimony analyses were done using software written by A.D.K. Once ancestral states were inferred, polarized mutations were classified as A/T to G/C or G/C to A/T. To explicitly deal with effects of polymorphism on estimates of ML divergence, we calculated distances for all possible permutations of unrooted, three-taxon trees (*D. melanogaster*, *D. simulans*, and *D. yakuba*) using one sampled allele for each species and reported the mean divergence along each branch.

Most statistical hypothesis testing was implemented using the R language (http://www.r-project.org). Stochasticity for base composition was assessed by comparing polymorphic versus fixed, A/T to G/C versus G/C to A/T mutations, following McDonald and Kreitman (1991). The nonparametric Wilcoxon rank sum test was used to test for significant differences in the frequency of segregating A/T to G/C and G/C to A/T polymorphisms.

Bayesian Analysis of Base-Composition Evolution

Under the assumption of selective neutrality, equilibrium base-composition bias can be estimated from present
base composition (A/T%) and estimates of the per bp rate of mutations towards (v) and away from (u) A/T (Sueoka 1993; Eyre-Walker 1997). Let

$$\omega = \frac{v}{u + v}$$

equal the mutation bias towards A/T. Eyre-Walker (1997) has shown that given a recent change in mutation bias, the expected equilibrium base-composition bias can be estimated by

$$\omega^* = \frac{fz}{1 - z - f + 2zf},$$

where $f$ is the current frequency of A/T nucleotides, and $z$ is the proportion of segregating mutations that are either A or T. One can use maximum likelihood to estimate $\omega$ under the assumption that the four types of observations in the analysis, sites that are fixed for an A/T, sites that are fixed for a G/C, and sites that are fixed for A/T→G/C and G/C→A/T polymorphisms, are multinomially distributed (Eyre-Walker 1997). Explicitly, if we denote the four types of observations to be $x_i$, $x_2$, $x_3$, and $x_4$, then the likelihood of observing the $x_i$ given the true proportions of sites are $y_i$, is given by

$$L = \left( \prod x_i \right)^{\frac{1}{y_i}} \prod x_i^{y_i}.$$
levels of intron polymorphism were correlated with levels of statistics of polymorphism and divergence. In general, looking at the statistic \( \sqrt{R} \) of \( h \), as suggested by Gelman et al. (1995). Based on this criterion, chains converged very quickly (< 4,000 iterations), as might be expected from such a simple likelihood function. Markov chains were thinned to sampling every 1,000 iterations, and 10,000 samples were taken after an initial burn-in of 5,000 iterations.

**Results**

**Polymorphism**

Tables 1 and 2 provide basic polymorphism and divergence statistics for intron sequences from *D. melanogaster* and *D. simulans*. Table 3 provides a polymorphism summary for intergenic loci from both *D. melanogaster* and *D. simulans*. Table 4 gives means across loci for summary statistics of polymorphism and divergence. In general, levels of intron polymorphism were correlated with levels of synonymous polymorphism from their corresponding genes in both *D. melanogaster* and *D. simulans* (*D. melanogaster*: R-squared = 0.5296, *P* < 0.0001; *D. simulans*: R-squared = 0.6033, *P* < 0.00001).

Mean intron heterozygosity was slightly lower than mean synonymous heterozygosity in both species (*D. melanogaster*: mean intron \( h_x = 0.0089 \), mean synonymous \( h_s = 0.0141 \); *D. simulans*: mean intron \( h_x = 0.0215 \), mean synonymous \( h_s = 0.0264 \)), although the mean differences were not statistically significant. This is consistent with an earlier analysis from a smaller data set (Moriyama and Powell 1996). However, a sign test on intron versus synonymous heterozygosity across genes was significant in *D. melanogaster* (18/24; *P* = 0.008) but not in *D. simulans* (15/24; *P* = 0.0779). Thus, *D. melanogaster* intron heterozygosity was consistently lower than the synonymous heterozygosity in the associated gene.

Levels of intergenic DNA polymorphism in *D. simulans* (mean \( h_x = 0.0102 \), median \( h_x = 0.0091 \)) were surprisingly low relative to intron and synonymous estimates. A Kruskal-Wallis rank sum test on heterozygosity for the three classes of silent polymorphisms in *D. simulans* (table 4) is nearly significant (\( \theta_x \): chi-square = 5.7789, *P* = 0.0553; \( \theta_y \): chi-square = 5.7924, *P* = 0.05523). In *D. melanogaster*, intergenic heterozygosity is roughly similar to intronic and synonymous heterozygosity (mean intergenic \( \theta_x = 0.0126 \); median intergenic \( \theta_x = 0.0126 \)). Kruskal-Wallis tests on the three categories of silent variation in *D. melanogaster* did not reject the null hypothesis of homogeneity (\( \theta_x \): chi-square = 4.228, *P* = 0.1211; \( \theta_y \): chi-square = 5.5167, *P* = 0.0634).

**D. melanogaster versus D. simulans heterozygosity**

Comparison of heterozygosities in *D. melanogaster* and *D. simulans* reveals significant differences for intronic
Table 3
Summary Statistics of DNA Polymorphism for Seven Intergenic Loci from D. melanogaster and D. simulans

<table>
<thead>
<tr>
<th>Species</th>
<th>Locus</th>
<th>$n$</th>
<th>bp $^b$</th>
<th>$S$ $^c$</th>
<th>$\theta_0$ $^d$</th>
<th>$\theta_e$</th>
<th>Tajima's $D$ $^e$</th>
<th>Indel Rate$^f$</th>
<th>J-C Div $^g$</th>
<th>Polarized Div $^h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td>inter11C1A</td>
<td>5</td>
<td>532.00</td>
<td>11</td>
<td>0.0102</td>
<td>0.0102</td>
<td>-0.0273</td>
<td>0.023</td>
<td>0.0401</td>
<td>0.0430</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter11C1A</td>
<td>8</td>
<td>119.75</td>
<td>3</td>
<td>0.0065</td>
<td>0.0101</td>
<td>-1.4475</td>
<td>0.008</td>
<td>0.0303</td>
<td>0.0239</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter21F</td>
<td>9</td>
<td>799.89</td>
<td>35</td>
<td>0.0149</td>
<td>0.0163</td>
<td>-0.4527</td>
<td>0.006</td>
<td>0.0564</td>
<td>0.0352</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter21F</td>
<td>8</td>
<td>640.38</td>
<td>24</td>
<td>0.0126</td>
<td>0.0154</td>
<td>-0.9561</td>
<td>0.008</td>
<td>0.0795</td>
<td>0.0551</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter49F</td>
<td>7</td>
<td>682.71</td>
<td>28</td>
<td>0.0148</td>
<td>0.0169</td>
<td>-0.7115</td>
<td>0.003</td>
<td>0.0610</td>
<td>0.0309</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter49F</td>
<td>8</td>
<td>928.13</td>
<td>15</td>
<td>0.0046</td>
<td>0.0063</td>
<td>-1.3893</td>
<td>0.001</td>
<td>0.0509</td>
<td>0.0809</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter86C</td>
<td>7</td>
<td>524.71</td>
<td>32</td>
<td>0.0249</td>
<td>0.0253</td>
<td>-0.0893</td>
<td>0.006</td>
<td>0.0861</td>
<td>0.0442</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter86C</td>
<td>8</td>
<td>528.50</td>
<td>2</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0000</td>
<td>0.0014</td>
<td>0.0041</td>
<td>0.0077</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter11C1B</td>
<td>8</td>
<td>118.63</td>
<td>1</td>
<td>0.0036</td>
<td>0.0033</td>
<td>0.0000</td>
<td>0.0033</td>
<td>0.0303</td>
<td>0.0120</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter11C1B</td>
<td>8</td>
<td>798.50</td>
<td>54</td>
<td>0.0290</td>
<td>0.0261</td>
<td>0.5902</td>
<td>0.0261</td>
<td>0.0564</td>
<td>0.0334</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter49F</td>
<td>7</td>
<td>689.57</td>
<td>26</td>
<td>0.0119</td>
<td>0.0159</td>
<td>-1.4484</td>
<td>0.0159</td>
<td>0.0610</td>
<td>0.0254</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter49F</td>
<td>8</td>
<td>663.38</td>
<td>24</td>
<td>0.0095</td>
<td>0.0142</td>
<td>-1.7466*</td>
<td>0.0142</td>
<td>0.0795</td>
<td>0.0468</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>inter84E</td>
<td>7</td>
<td>902.71</td>
<td>20</td>
<td>0.0097</td>
<td>0.0091</td>
<td>0.0187</td>
<td>0.0091</td>
<td>0.0509</td>
<td>0.0469</td>
</tr>
<tr>
<td>D. simulans</td>
<td>inter84E</td>
<td>8</td>
<td>564.50</td>
<td>12</td>
<td>0.0070</td>
<td>0.0083</td>
<td>-0.7605</td>
<td>0.0083</td>
<td>0.0861</td>
<td>0.0387</td>
</tr>
</tbody>
</table>

$^a$ Species and Locus are the species and loci surveyed.
$^b$ n is the number of alleles sampled.
$^c$ S is the number of segregating sites in this sample, including those sites that contain ambiguous bases.
$^d$ $\theta_0$ and $\theta_e$ are estimates of nucleotide polymorphism; sites containing ambiguous bases have been removed for these estimates.
$^e$ Tajima's $D$ is a measure of the site frequency spectrum.
$^f$ Indel rate is the number of unique, nonoverlapping indels per site.
$^g$ J-C Div is the Jones-Cantor corrected divergence between D. melanogaster and D. simulans.
$^h$ Polarized Div is the average ML divergence along a lineage over all possible unrooted three-taxon trees.

and synonymous (Wilcoxon rank sum tests, intronic $\theta_0$: $P = 0.0006$; synonymous $\theta_e$: $P = 0.0203$) but not for intergenic sites (Wilcoxon rank sum tests, intergenic $\theta_0$: $P = 0.3176$). However, comparison of African D. melanogaster samples (n = 11 loci: mth, G6pd, mei-218, pgm, tpi, Adh, anon1A3, anon1E9, anon1G5, v, and acp53 [table 1 in Supplementary Material online]) versus D. simulans provides a slightly different result. Intron variation is significantly lower in D. melanogaster (Wilcoxon rank sum test, intronic $\theta_0$: $P = 0.0027$), but intergenic and synonymous heterozygosity are roughly equal (African D. melanogaster: mean intergenic $\theta_0$ = 0.0126, mean intronic $\theta_0$ = 0.0097; median synonymous $\theta_0$ = 0.0176; D. simulans: mean intergenic $\theta_0$ = 0.0102, mean intronic $\theta_0$ = 0.0205, mean synonymous $\theta_0$ = 0.0244). In other words, for African D. melanogaster versus D. simulans, the only obvious difference in heterozygosity is for introns.

Divergence

Unpolarized comparisons of divergence between D. melanogaster and D. simulans among silent sites reveals that the rank order of divergence is intergenic (mean Jukes-Cantor corrected divergence = 0.0578) < in intronic (mean JC corrected = 0.0969) < synonymous (mean JC corrected = 0.1181). Thus, in pairwise comparisons, intergenic DNA is evolving roughly half as quickly as intronic or synonymous sites. A test for heterogeneity of divergences among classes of sites is highly significant (Kruskal-Wallis rank sum test; chi-square = 14,7973, $P = 0.0006$). Mean divergence for the two classes of non-coding sites (i.e., intergenic and intronic sequence) is also significantly different (Wilcoxon rank sum test, $P = 0.0291$).

Divergence that has been polarized to the D. melanogaster or D. simulans lineages (see Materials and Methods) provides a similar picture of evolution. Rank order mean divergence in D. melanogaster for silent classes was intergenic (0.0447) < synonymous (0.0571) < in intronic (0.0730). These three classes are not significantly heterogeneous (Kruskal-Wallis rank sum test, chi-square = 2.3284, $P = 0.3122$). This rank order of mean divergence is slightly different from that observed in D. simulans, which mirrors the unpolarized analysis: intergenic (0.0301) < in intronic (0.0451) < synonymous (0.0506). The three site classes were also not heterogeneous for D. simulans (Kruskal-Wallis rank sum test, chi-square = 4.2384, $P = 0.1203$).

Our analysis supports the idea that D. melanogaster evolves faster than D. simulans across all silent sites (Wilcoxon rank sum test, silent sites $P = 0.0202$; sign test, 38/56 $P = 0.0029$). This is consistent with previous reports from protein-coding regions (Akashi 1996).

Tests of Neutrality

We examined the frequency spectrum of segregating sites and compared levels of polymorphism and divergence to test for departures from neutral equilibrium expectations at silent sites. Tests of the frequency spectrum

Table 4
Mean Polymorphism and Divergence Summary

<table>
<thead>
<tr>
<th>Silent Site type</th>
<th>Summary Statistic</th>
<th>D. melanogaster</th>
<th>D. simulans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic $\theta$</td>
<td>0.0089</td>
<td>0.0215</td>
<td></td>
</tr>
<tr>
<td>Polarized divergence</td>
<td>0.0730</td>
<td>0.0451</td>
<td></td>
</tr>
<tr>
<td>Intergenic $\theta$</td>
<td>0.0126</td>
<td>0.0102</td>
<td></td>
</tr>
<tr>
<td>Polarized divergence</td>
<td>0.0447</td>
<td>0.0301</td>
<td></td>
</tr>
<tr>
<td>Synonymous $\theta$</td>
<td>0.0141</td>
<td>0.0264</td>
<td></td>
</tr>
<tr>
<td>Polarized divergence</td>
<td>0.0571</td>
<td>0.0506</td>
<td></td>
</tr>
</tbody>
</table>

Note.—See table 1 and Materials and Methods for explanation of summary statistics.
were carried out using Tajima’s $D$ statistic (Tajima 1989). The results from these tests can be found in tables 1, 2, and 3. Generally, there were few instances of significant $D$ values. We carried out multilocus HKA tests of synonymous, intronic, and intergenic sites separately. Figures 1, 2, and 3 display locus-by-locus contributions to the overall test statistic. Observed values that were less than/greater than expected values are plotted as negative/positive values, respectively. Both synonymous and intronic sequences (figs. 1 and 2) yielded a significant HKA result (intronic: $P = 0.0170$; synonymous: $P = 0.0002$), consistent with too little polymorphism, given levels of divergence. Deviations for $D. melanogaster$ polymorphism, $D. simulans$ polymorphism, and divergence between species were highly correlated among intronic and synonymous sites from a given locus (Spearman’s rank correlation: $D. melanogaster$ polymorphism, $rho = 0.4617$, $P = 0.02412$; $D. simulans$ polymorphism, $rho = 0.4497$, $P = 0.02851$; interspecific divergence, $rho = 0.5904$, $P = 0.002837$). To the extent that the significant heterogeneity as revealed by HKA tests reflects the impact of linked, beneficial mutations (Maynard Smith and Haigh 1974; Kaplan et al. 1989), the correlations of deviations suggest that hitchhiking effects often span gene-sized or larger genomic regions. Intergenic sequences, however, did not reject neutrality ($P < 0.6920$). The number of intergenic regions surveyed ($n = 7$) is much smaller that the number of introns surveyed ($n = 49$). Thus, the different statistical results for intron + synonymous versus intergenic could reflect reduced power to reject the null in the intergenic sample or could result if most beneficial mutations occur in or near genes, rather than in intergenic DNA. However, if we restrict our HKA tests of synonymous and intronic sites to African $D. melanogaster$ samples, we observe a slightly different result. In this case, synonymous sites still reject neutrality, but intronic sites do not reject the null (African samples, synonymous: $P = 0.0430$; intronic: $P < 0.389$). Once again it is feasible that this difference might be the result of reduced power to reject in our smaller sample of African loci/alleles. Alternatively, selection associated with migration out of Africa might
explain the differences seen between our worldwide samples and our restricted African subsamples (e.g., Kauer, Dieringer, and Schlotterer 2003).

Base-Composition Evolution

We used ancestral reconstructions to polarize G/C to A/T and A/T to G/C fixations and polymorphisms for the *D. melanogaster* and *D. simulans* lineages, respectively. As few mutations of any class were observed within individual loci, we combined data from across loci to make a more general statement of genomic evolution. Tables 5, 6, and 7 present these aggregate data in the form of a 2 × 2 contingency table for each species, for intronic, intergenic, and synonymous sites, respectively. Interestingly, if we restrict our attention to only African *D. melanogaster* samples, all of our base-composition results remain qualitatively unchanged.

Substitutions

Sequences evolving at equilibrium for base composition are expected to fix equal numbers of G/C to A/T and A/T to G/C mutations. Intron and intergenic fixations are consistent with equilibrium for both species (*D. melanogaster* binomial probability: introns ML \( P = 0.523 \), parsimony \( P = 0.5145 \); intergenic ML \( P = 0.6254 \), parsimony \( P = 1.0 \) and *D. simulans* binomial probability: introns ML \( P = 1.0 \), parsimony \( P = 0.435 \); intergenic ML \( P = 0.5114 \), parsimony \( P = 0.7283 \)). However, *D. melanogaster* and *D. simulans* synonymous sites are fixing significantly more A/T than G/C mutations (*D. melanogaster* binomial probability: synonymous sites ML \( P > 1 \times 10^{-17} \), parsimony \( P < 1 \times 10^{-22} \) and *D. simulans* binomial probability: synonymous sites ML \( P = 0.0441 \), parsimony \( P = 0.0403 \)), a pattern which has been noted before (Akashi 1994; Begun 2001).

Contrasts of Polymorphism and Divergence

Comparisons of polymorphic and fixed G/C to A/T and A/T to G/C mutations reveal striking patterns of heterogeneity. Although the 2 × 2 table of polymorphic and fixed mutations is not significantly heterogeneous for *D. simulans* introns, *D. simulans* intergenic, or *D. melanogaster* intergenic DNA, the table for *D. melanogaster* introns is significantly heterogeneous, regardless of the method of ancestral reconstruction used (G-tests: ML \( P = 0.0013 \), parsimony \( P = 0.008 \)). It appears that there is a large excess of G/C to A/T polymorphisms in *D. melanogaster* introns.

Table 5

**Intron Base-Composition Evolution**

<table>
<thead>
<tr>
<th>Species</th>
<th>State</th>
<th>( \text{A/T to G/C} )</th>
<th>( \text{G/C to A/T} )</th>
<th>( \text{A/T to G/C} )</th>
<th>( \text{G/C to A/T} )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. melanogaster</em></td>
<td>Fixed</td>
<td>56</td>
<td>64</td>
<td>39</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polymorphic</td>
<td>15</td>
<td>50</td>
<td>12</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( P = 0.0013^{c} )</td>
<td></td>
<td>( P = 0.0080^{c} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. simulans</em></td>
<td>Fixed</td>
<td>21</td>
<td>20</td>
<td>33</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polymorphic</td>
<td>47</td>
<td>46</td>
<td>42</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( P = 0.942^{c} )</td>
<td></td>
<td>( P = 0.441^{c} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\) ML values are based upon maximum-likelihood ancestral reconstruction under the GTR model and subsequent polarization of fixations and polymorphisms.

\(^{b}\) Parsimony is the same as above, except which ancestral reconstruction is done in a maximum-parsimony framework.

\(^{c}\) \( P \) values are from log-likelihood ratio goodness-of-fit tests.
The 2 × 2 tables for synonymous site evolution (table 7) are highly significant in *D. simulans* and *D. melanogaster*. Previous analyses of *D. simulans* variation revealed a major excess of A/T polymorphisms and smaller excess of A/T fixations; the polymorphism and divergence was heterogeneous in that the deviations from equilibrium expectation were much greater for the polymorphism (Akashi 1996; Begun 2001). Our results are consistent with these earlier analyses. Surprisingly, however, the 2 × 2 tables for synonymous site evolution (table 7) show that *D. melanogaster* is also significantly heterogeneous for polymorphic and fixed G/C versus A/T variants. Previous analyses of fewer *D. melanogaster* data showed no heterogeneity of polymorphic versus fixed A/T versus G/C mutations at synonymous sites (Akashi 1994, 1995).

**Frequency Distribution**

Given the above results from *D. melanogaster*, we investigated the frequency distribution of A/T and G/C polymorphisms from intronic and intergenic DNA in *D. melanogaster*. Comparisons using different ancestral reconstructions yield inconsistent results in this analysis. If we consider each site as an independent observation and use parsimony reconstructions, G/C mutations are at higher frequency than A/T mutations (Wilcoxon rank sum test: *P* = 0.03131). ML reconstructions show a qualitatively similar pattern; however, a hypothesis test fails to reject the null (Wilcoxon rank sum test: *P* = 0.1701). Thus, there is weak evidence that derived G/C and A/T mutations are at different frequencies in *D. melanogaster* populations. If we compare the average frequency per gene among silent classes (a more conservative test), we do not detect a significant difference between classes. Intergenic regions from *D. melanogaster* also show no hint of a frequency difference among polymorphism classes, both when comparing gene-averaged frequencies, or when treating segregating sites as independent observations. It is worth noting, however, that failure to reject the null in this setting may be the result of reduced power in the small population samples we have examined.

**Bayesian Analysis of Base-Composition Evolution**

Under the premise that excess A/T intron polymorphism in *D. melanogaster* results from a recent change in mutation bias, we can estimate the new expected mutation bias at equilibrium *α* (Eyre-Walker 1997). Figure 4 presents the posterior distribution of *α*, given the numbers of A/T and G/C polymorphisms and fixations from *D. melanogaster* and *D. simulans* introns from our Markov chain simulation. Table 8 presents results from our MCMC contrasting *D. melanogaster* and *D. simulans* at both intronic and intergenic sites. Current A/T content in *D. melanogaster* introns is 0.647. The MAP estimate of *α* from our MCMC simulation was 0.77, with a 95% credible interval from 0.651 to 0.846. Thus, the data are compatible

---

**Table 6**

<table>
<thead>
<tr>
<th>Intergenic Base-Composition Evolution</th>
<th>MLa</th>
<th>Parsimonyb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>48</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td><em>P</em> = 0.6899c</td>
<td><em>P</em> = 0.6182c</td>
</tr>
<tr>
<td><em>D. simulans</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td><em>P</em> = 0.6279c</td>
<td><em>P</em> = 0.3981c</td>
</tr>
</tbody>
</table>

* a ML values are based upon maximum-likelihood ancestral reconstruction under the GTR model and subsequent polarization of fixations and polymorphisms.

b Parsimony is the same as above, except which ancestral reconstruction is done in a maximum-parsimony framework.

c *P*-values are from log-likelihood ratio goodness-of-fit tests.

**Table 7**

<table>
<thead>
<tr>
<th>Synonymous Site Base-Composition Evolution</th>
<th>MLa</th>
<th>Parsimonyb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>80</td>
<td>38</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td><em>P</em> &lt; 0.0001c</td>
<td><em>P</em> &lt; 0.001c</td>
</tr>
<tr>
<td><em>D. simulans</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>53</td>
<td>64</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>104</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td><em>P</em> = 0.0036c</td>
<td><em>P</em> = 0.0018c</td>
</tr>
</tbody>
</table>

* a ML values are based upon maximum-likelihood ancestral reconstruction under the Goldman-Yang codon model and subsequent polarization of fixations and polymorphisms.

b Parsimony is the same as above, except which ancestral reconstruction is done in a maximum-parsimony framework.

c *P*-values are from log-likelihood ratio goodness-of-fit tests.
with a major evolutionary change of base composition in
D. melanogaster introns. This is in contrast to our esti-
mates from D. simulans introns, where the current A/T
frequency, 0.614, is well within the 95% credible interval
for our estimates of \( \omega \) (table 8). Additionally, MCMC
simulations for our intergenic data are compatible, with no
change in equilibrium base-composition bias, over that
currently observed in either species.

Discussion

Though D. melanogaster and D. simulans have been
important model systems in molecular population genetics
for several years, most molecular population genetic data
from these species have been from protein-coding regions.
Our analysis of noncoding DNA revealed two properties
of the substitution process.

First, intergenic DNA appears to evolve more slowly
than intron DNA or synonymous sites in exons. This is
consistent with recent reports by Halligan et al. (2004) on
intergenic divergence in these species. However, much of
their intergenic sequence was sampled from 5' and 3'
flanking regions of genes. Such regions may be expected
to harbor functionally important sites. Conversely, much
of our sampled intergenic DNA comes from regions not
near known or predicted genes. One interpretation of our
data (and those of Halligan et al. [2004]) is that the neutral
mutation rate in intergenic DNA is unexpectedly low as
a result of functional constraint. The relatively low
heterozygosity in these regions for D. simulans is
consistent with this interpretation. However, the data are
also completely consistent with a reduced mutation rate in
nontranscribed sequences in Drosophila.

Second, D. melanogaster appears to be evolving
significantly faster than D. simulans at all three classes of
silent sites: synonymous, intergenic, and intronic. Previous
results (Akashi 1996) suggested that rates of protein
evolution are higher along the D. melanogaster than the D.
simulans lineage. Indeed, analysis of the larger set of genes
used here confirms the trend of faster protein evolution
along the D. melanogaster lineage, although the difference
is not significant (D. melanogaster: mean \( d_{N} \) = 0.0131; D.
simulans: mean \( d_{N} \) = 0.0124). Thus, the data from D.
melanogaster and D. simulans suggest that D. mel-
onganaster evolves faster than D. simulans at all classes of
sites examined. Akashi (1996) interpreted the higher rate
of evolution in D. melanogaster replacement and synon-
ymous sites (versus D. simulans) as resulting from reduced
efficacy of natural selection against slightly deleterious
mutations, perhaps as a consequence of reduced effective
population size. If this explanation is correct, then our data
suggest that all classes of sites are fixing slowly
deleterious mutations in the D. melanogaster lineage.
A corollary of this hypothesis is that the distribution of
selection coefficients of new mutations is similar (i.e.,
nearly neutral) across these functionally disparate site
types. Perhaps a more parsimonious interpretation is that
the mutation rate has increased in the D. melanogaster
lineage (relative to D. simulans). This would seem to be at
odds with the observation that D. melanogaster is less
polymorphic than D. simulans in both African and non-
African populations (e.g., Begun and Whitley 2000;
Andolfatto 2001). However, the neutral model posits that
levels of variation depend on population size and neutral
mutation rate, whereas divergence only depends on neutral
mutation rate (Kimura 1983) Perhaps one attractive feature
of the mutation rate–increase hypothesis is that it is
amenable to direct empirical testing (e.g., Eeken, de Jong,
and Green 1987).

Previous studies of D. melanogaster and D. simulans
variation suggested a fundamental difference between
species in the dynamics of synonymous variation.
Contrasts of polymorphic versus fixed, A/T versus G/C
mutations in D. melanogaster were homogeneous, yet A/T

![Posterior Distribution of Bias](image)

**FIG. 4.** —The posterior distribution of \( \omega \), the expected equilibrium base-composition bias. The distribution is derived from four chains of 10,000 iterations each. Current A/T content in D. melanogaster introns is 0.647. The MAP estimate of \( \omega \) from our simulations was 0.77, with a 95% credible interval from 0.651 to 0.846.

### Table 8

Bayesian Analysis of Expected Equilibrium Base-Composition Bias

<table>
<thead>
<tr>
<th>Silent Site Type</th>
<th>Species</th>
<th>Current A/T%a</th>
<th>MAP Estimate of ( \omega )b</th>
<th>95% Credible Interval of ( \omega )c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introns</td>
<td>D. melanogaster</td>
<td>0.626</td>
<td>0.770</td>
<td>0.651–0.846</td>
</tr>
<tr>
<td></td>
<td>D. simulans</td>
<td>0.614</td>
<td>0.485</td>
<td>0.307–0.659</td>
</tr>
<tr>
<td>Intergenic</td>
<td>D. melanogaster</td>
<td>0.634</td>
<td>0.510</td>
<td>0.379–0.681</td>
</tr>
<tr>
<td></td>
<td>D. simulans</td>
<td>0.631</td>
<td>0.570</td>
<td>0.402–0.752</td>
</tr>
</tbody>
</table>

a Current A/T% was estimated as the average frequency across alleles where polymorphism data were available.
b Maximum a posteriori (MAP) estimate of \( \omega \) from the Bayesian estimation procedure described in the methods section.
c The 95% credible interval of \( \omega \) from the Bayesian estimation procedure described in the Materials and Methods section.
mutations greatly outnumbered G/C mutations for both polymorphic and fixed sites (Akashi 1995). The contrast of polymorphic versus fixed, A/T versus G/C mutations in D. simulans was highly significantly heterogeneous, consistent with an excess of A/T polymorphisms (Akashi 1995; Begun 2001). These results were interpreted as supporting a model of neutral evolution of higher A/T content in D. melanogaster (Akashi 1995, 1996). D. simulans data suggested that this lineage is also evolving higher A/T content (Begun 2001, McVean and Vieira 2001), although at a much slower rate than D. melanogaster, and that the polymorphism may be inconsistent with neutrality (Akashi 1996).

Our analysis of this larger D. melanogaster set suggests a different view of the polymorphism and divergence at synonymous sites, namely, that the contrasts of polymorphic and fixed A/T versus G/C mutations are significantly heterogeneous in both species (table 7). Such allele configurations (table 7) were previously interpreted as evidence for reduced selection against slightly deleterious A/T fixations after a population bottleneck and as evidence for nearly neutral dynamics for A/T polymorphisms (Akashi 1995, 1996). An alternative hypothesis for such synonymous site data is a recent change in mutation bias (Eyre-Walker 1997; Akashi 1997). Of course, rejection of homogeneity in the 2 × 2 contingency table is formally consistent with other interpretations, which complicates inferences regarding base-composition evolution at synonymous sites. For example, if the polymorphism data provide an accurate picture of historical mutational input, the data could be interpreted as a lineage-specific fixation bias of G/C mutations in D. melanogaster, perhaps as result of natural selection or biased gene conversion (Holmiquist 1992; Eyre-Walker 1993, 1999). In any case, our results suggest that patterns of base-composition variation at synonymous sites are qualitatively similar in the two species.

If the mutation bias hypothesis was the correct explanation for synonymous site variation, we might expect to observe similar patterns of base-composition evolution for intronic and intergenic sites. In fact, we did observe a similar pattern, but only for intronic and not intergenic sites, and only for D. melanogaster and not D. simulans (tables 5 and 6). Thus, the data from these two species are recalcitrant to a simple explanation. An analysis aimed at quantifying the magnitude of such a bias from D. melanogaster introns is consistent with a large change in base composition currently under way along that lineage (fig. 4).

One interpretation of the lineage and site-specific heterogeneity for contrasts of polymorphic and fixed, A/T versus G/C mutations is a recent change in mutation bias in transcribed D. melanogaster (but not in D. simulans) sites (i.e., introns and synonymous sites). In principle, this could explain both the much greater excess of A/T fixations in D. melanogaster versus D. simulans synonymous sites and the excess of A/T polymorphisms in D. melanogaster but not in D. simulans introns. Overall then, data from the two species may be consistent with mildly deleterious evolution of synonymous sites in both species, leading to an accumulation of unpreferred (i.e., A/T ending) codons and a recent change in mutation bias associated with transcribed DNA in D. melanogaster. This explanation is somewhat unsatisfying because it is essentially a restatement of the empirical observations. However, it does have the virtue of predicting that careful measurement of DNA metabolism in D. melanogaster and D. simulans will reveal interspecific differences. For example, the hypothesis of different patterns of mutation bias in D. melanogaster versus D. simulans could be investigated experimentally by direct assays of genomic DNA damage or genetic assays to detect differences in DNA repair efficiency between D. melanogaster and D. simulans. Genomic-scale polymorphism and divergence data from DNA sequences varying in levels of germline transcription would also be valuable for further exploration of these ideas.

Mammalian data suggest that germline transcription-coupled repair may have important consequences for base-composition evolution (Green et al. 2003, Majewski 2003). However, there is no evidence for transcription-coupled repair in flies (de Cock et al. 1992; van der Helm et al. 1997; Sekelsky, Brodsky, and Burris 2000). Nevertheless, transcribed DNA is packaged into a more “open” chromatin conformation than nontranscribed DNA (reviewed in Kornberg and Lorch [1995]), which is thought to make transcribed DNA more accessible to endogenous DNA-damaging agents (Bartlett, Scicchitano, and Robison 1991; Ljungman and Hanawalt 1992). In principle, differential chromatin states may affect the distribution of new mutations in transcribed versus nontranscribed sequences—a phenomenon known as transcription-associated mutation (TAM).

A problem with the recent change in mutation-bias hypothesis is that it predicts G/C to A/T mutations are on average, younger than A/T to G/C mutations and, thus, should be segregating at lower frequencies in D. melanogaster (Sawyer 1977). We observed no such heterogeneity of frequencies, although the relatively small sample sizes may provide little power to detect such a phenomenon.

A recent survey of synonymous sites at the rosy locus from 22 Drosophila species showed that many lineages appear to be evolving higher A/T content (Begun and Whitley 2003). In this respect, D. melanogaster and D. simulans appear to be like other flies—they are becoming more A/T rich at synonymous sites. One explanation for these data is that the ancestor of Drosophila was highly G/C rich but evolved a strong(er) A/T mutation bias before splitting of the extant Drosophila lineages. In this scenario, the accumulation of A/T mutations in many lineages reflects a slow approach to a new, higher A/T–content equilibrium. If intergenic and intron sequences are under weaker selection for base composition, perhaps they would have approached this equilibrium more quickly compared with synonymous sites, thereby explaining why the substitution data from intergenic and intron DNA, but not synonymous sites, supports base-composition equilibrium.

Finally, one of the major questions of population genetics is how much of the genome is either directly or indirectly influenced by directional selection. Our comparisons of polymorphism and divergence from worldwide samples using the HKA test (Hudson, Kreitman and
Aguadé (1987) revealed that whereas intergenic data were consistent with neutrality, synonymous site and intronic data were not (although note the differences in our restricted African sample). If the significant HKA results are a consequence of hitchhiking effects (Maynard Smith and Haigh 1974; Kaplan et al. 1989), then we might infer that (1) directional selection tends to occur in genes and (2) the physical extent of hitchhiking effects is often “gene-sized” or larger—on the order of a few kilobases.

Acknowledgments

We thank John Gillespie, Chuck Langley, Matt Hahn, Corbin Jones, Josh Chang Mell, and the UC Davis Evolutionary Discussion Group (EDG) for valuable comments and discussion. Michael Nachman and two anonymous reviewers improved this manuscript substantially. A.D.K. is a Howard Hughes Medical Institute predoctoral fellow. This work was supported by NSF DEB-0327049. M. Kerber provided technical assistance.

Literature Cited


Bartlett, J. D., A., Scicchitano, and S. H. Robison. 1991. Two anonymous reviewers improved this manuscript substantially. A.D.K. is a Howard Hughes Medical Institute predoctoral fellow. This work was supported by NSF DEB-0327049. M. Kerber provided technical assistance.


Michael Nachman, Associate Editor

Accepted August 30, 2004