DOI: 10.1534/genetics.104.032250

Genomic Heterogeneity of Background Substitutional Patterns in *Drosophila melanogaster*

Nadia D. Singh,*^,1 Peter F. Arndt† and Dmitri A. Petrov*

**Department of Biological Sciences, Stanford University, Stanford, California 94305*
and †Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

Manuscript received June 15, 2004
Accepted for publication October 28, 2004

ABSTRACT

Mutation is the underlying force that provides the variation upon which evolutionary forces can act. It is important to understand how mutation rates vary within genomes and how the probabilities of fixation of new mutations vary as well. If substitutional processes across the genome are heterogeneous, then examining patterns of coding sequence evolution without taking these underlying variations into account may be misleading. Here we present the first rigorous test of substitution rate heterogeneity in the *Drosophila melanogaster* genome using almost 1500 nonfunctional fragments of the transposable element DNAREP1_DM. Not only do our analyses suggest that substitutional patterns in heterochromatic and euchromatic sequences are different, but also they provide support in favor of a recombination-associated substitutional bias toward G and C in this species. The magnitude of this bias is entirely sufficient to explain recombination-associated patterns of codon usage on the autosomes of the *D. melanogaster* genome. We also document a bias toward lower GC content in the pattern of small insertions and deletions (indels). In addition, the GC content of noncoding DNA in Drosophila is higher than would be predicted on the basis of the pattern of nucleotide substitutions and small indels. However, we argue that the fast turnover of noncoding sequences in Drosophila makes it difficult to assess the importance of the GC biases in nucleotide substitutions and small indels in shaping the base composition of noncoding sequences.

MUTATION is the substrate of evolution; it creates the variation upon which evolutionary forces will ultimately act. With every novel mutation there is an associated probability of fixation, which may be affected by neutral forces, such as random genetic drift or biased gene conversion, or selective forces. Both mutation rates and fixation probabilities of new mutations may vary across the genome, which may play a role in generating patterns of extant sequence variation. One of the goals of genomic biology is to understand how heterogeneity in evolutionary forces such as mutation, drift, biased gene conversion, and selection can actively shape genome sequences and structure.

The evolution of both functional and nonfunctional sequences might be modulated by any or all of these forces, though perhaps to different extents. It is likely that mutational patterns and rates are going to be comparable in both functional and nonfunctional regions of the genome. It is also probable that certain fixation biases, for example, due to biased gene conversion or selection at the level of global GC content, will affect functional and nonfunctional sequences in a similar way. However, in addition to these background patterns of substitutions, functional sequences are likely to have unique fixation biases driven by selective forces specific to their functional capacity. In some cases, such selective effects are well known, such as those related to maintenance of the integrity of protein-coding sequences, but many other, possibly more subtle, patterns are likely yet to be discovered.

In this context, the study of the background patterns of substitution not only is interesting in its own right, but also should help us rigorously define null hypotheses against which we can compare patterns of substitution found in functional sequences such as genes. This, in turn, should help us identify selective forces that act within genes and thereby understand which genic features are of functional significance.

In this genomic age, the need for careful quantification of background substitutional patterns has become particularly acute. More and more studies compare patterns of evolution among genes scattered across the genome. Differences among these genes in nucleotide or protein evolution are often interpreted as indications of differences in gene function or in the strength of natural selection acting on functions encoded in genic sequences. However, systematic differences in background substitutional patterns among genomic regions may also exist. The background rates and patterns of substitutions may be different between euchromatic and heterochromatic sequences, for instance, or among genomic regions of varying GC content. Not taking these potential differences into account may introduce sys-

[1] *Corresponding author:* Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 94305-5020.
E-mail: ndsingh@stanford.edu

tematic errors into our understanding of the evolution of coding sequences.

Studying molecular evolution across gradients in recombination rate provides a particularly challenging and illustrative example. In many genomes recombination rate varies substantially (Nachman 2002; McVean *et al.* 2004). There are theoretical reasons to believe (Hill and Robertson 1966) and empirical data to suggest (Begun and Aquadro 1992; Aquadro 1997) that genes located in regions of reduced recombination are subject to weaker purifying selection and less effective positive selection as compared to genes in regions of high recombination. In regions of low recombination, multiple selected variants should interfere with one another's population dynamics, effectively making population processes more stochastic. This increased stochasticity is akin to a virtual decrease of the effective population size in regions of low recombination, which makes it very tempting to study molecular evolution along a recombination gradient within a single species. In using recombination rate as a proxy for effective population size, we can examine the effects of effective population size on rates of molecular evolution while still effectively controlling for population and evolutionary history in a way that would be virtually impossible in a comparison of two different species.

However, it is not immediately clear that background patterns of substitution are independent of recombination rate. For instance, recombination itself may be mutagenic. Not only would this increase the rate of substitution in sequences located in regions of high recombination, but also if recombination-associated mutations were biased, this could change the pattern of substitutions among different recombination regimes. Moreover, higher rates of recombination may lead to higher rates of gene conversion. To the extent that gene conversion is biased, this could systematically alter fixation probabilities of some mutations with respect to recombination rate. As a result, it is no longer sufficient to simply correlate changes in genic evolution with recombination rate and ascribe functional significance to those changes. Rather, to be rigorous in our treatment of coding sequence evolution, we must quantify variation in background substitutional patterns associated with recombination and test for deviations from those patterns to identify features of genic sequences that are uniquely under selection for function.

To quantify background rates and patterns of nucleotide substitutions, we need a collection of fairly similar sequences scattered across the genome that are unlikely to be participating intimately and essentially in gene function. It is also preferable to know in each case the ancestral sequence so that we can infer the number and direction of evolutionary changes. In large genomes we can often use pseudogenes or inactive copies of particular families of transposable elements to accomplish this goal. Indeed, the truly vast numbers of inactive transposable elements in the human genome have allowed a very precise spatial (at a 1-Mbp resolution) and temporal characterization (over the last 250 Myr) of the patterns of background substitution through the comparative analysis of multiple paralogous copies of inactive transposable elements (Arndt *et al.* 2003b). Unfortunately, in species with smaller genomes, such as in Drosophila, pseudogenes and inactive copies of transposable elements are generally rare, most likely because they are lost very rapidly through frequent small deletions (Petrov *et al.* 1996; Petrov 2002).

To circumvent this problem, one could analyze multiple families of transposable elements (Blumenstiel *et al.* 2002). While overall background substitutional biases can be investigated using this approach, there is a possibility that different transposable elements evolve in slightly different ways. Moreover, because different families of transposable elements have different transpositional histories, with elements transposing at varied times on an evolutionary timescale, it is difficult to ascertain whether a particularly diverged copy of a transposable element is located in a region of elevated background substitution rate or whether it simply inserted prior to other copies. Accordingly, an aggregate analysis of background rates of nucleotide substitutions across different families of transposable elements is difficult.

A recent examination of the *Drosophila melanogaster* genome revealed a DNA element, DNAREP1_DM, that is present at an unusually high copy number for this species (Kapitonov and Jurka 2003). This element appears to be a nonautonomous derivative of a *Penelope*-like element and there are several thousand copies of DNAREP1_DM scattered across the genome (Kapitonov and Jurka 1999, 2003). In addition, previous analyses of this element (Kapitonov and Jurka 1999, 2003; Singh and Petrov 2004) revealed that DNAREP1_DM was mobilized in a burst of transposition ~5 MYA and has remained inactive ever since. As a result, examination of this element allows the first comprehensive analysis of background rates and patterns of nucleotide substitutions in *D. melanogaster*.

We used this unique (for Drosophila) data set of DNAREP1_DM transposable element fragments to test for variation in background substitution rates associated with recombination on the autosomes of *D. melanogaster*. Our results revealed systematic variation in background substitution patterns between heterochromatic and euchromatic sequences and provided evidence that background substitutional patterns are biased toward G's and C's in regions of higher recombination. We also show that the GC content of noncoding sequences within the *D. melanogaster* genome is not in equilibrium with respect to the pattern of single-nucleotide substitutions, as current GC content appears to be consistently higher than expected under patterns of background substitution.

Having quantified variation in background patterns of substitution associated with recombination rate, we investigated whether the positive correlation between codon bias and recombination rate unique to the autosomes of *D. melanogaster* (N. D. Singh, J. C. Davis and D. A. Petrov, unpublished data) was better explained by the Hill-Robertson effect of greater efficacy of natural selection in regions of high recombination (Kliman and Hey 1993; Comeron *et al.* 1999; Hey and Kliman 2002) or by the systematic variation of background substitutional patterns. On the basis of our analysis, we cannot reject the null hypothesis that the correlation between codon bias and recombination rate on the autosomes of *D. melanogaster* is entirely due to background substitutional patterns, at least for several amino acids. These results do not rule out the possibility that Hill-Robertson interference governs recombination-associated patterns of codon usage for other amino acids within these same genes or within subsets of genes in the *D. melanogaster* genome (Marais and Piganeau 2002). Our results do not suggest that codon bias in *D. melanogaster* is not shaped by selection on translational efficiency but instead suggest that to the extent that selection shapes codon frequencies in *D. melanogaster*, the efficacy of this selection is not noticeably reduced in areas of low recombination.

MATERIALS AND METHODS

Sequence retrieval and partitioning: Extant copies of the nonautonomous transposable element DNAREP1_DM were retrieved from the *D. melanogaster* genome using default parameters in the standalone implementation of BLASTN. The query sequence for our BLAST search was the reported 594-bp consensus sequence for this element (Kapitonov and Jurka 1999), and this was queried against the individual chromosomal arms of the autosomes of the *D. melanogaster* genome (FlyBase version 3.2) with a [5, −4] scoring matrix. All BLAST hits not located in telomeric regions and >100 bp were included in our data set, and overlapping hits were merged. BLAST hits that were found in contigs 2h and 3h were designated as heterochromatic elements, while the remaining hits were classified as euchromatic.

Euchromatic copies of DNAREP1_DM were partitioned further into six categories: intron, 5′-UTR, 3′-UTR, upstream, downstream, and remaining, describing hits that fell within introns, 5′-untranslated regions, 3′-untranslated regions, within 1000 bp upstream of a gene, within 1000 bp downstream of a gene, and all remaining hits, respectively. Within each of these categories, copies of DNAREP1_DM were separated into two recombination categories, low and high, based on calculations described below. For each of our categories (heterochromatin, intron low recombination, etc.) all extant fragments of this element were concatenated, as were the ancestral sequences for those fragments. The concatenated daughter sequence was compared to the concatenated ancestral sequence to calculate rates of nucleotide substitution.

Estimation of substitution frequencies: We first confirmed that all of the bases of the transposable element DNAREP1_DM were roughly equally diverged from their consensus in all observed copies of the element. We defined divergence of an individual base as the number of observed changes at this position normalized by the number of instances this position is found in all the retrieved fragments. To test if the divergence is homogeneous along the element, we partitioned the consensus sequence into four equal-length segments and tested whether the distributions of divergence for these segments had different means in pairwise comparisons. After accounting for bases that were not found in any of our extant fragments and bases that did not vary, we detected no significant pairwise differences in mean divergence among regions after suitable correction for multiple testing. Therefore, the assumption that there are no regional differences within the element DNAREP1_DM with respect to their divergence seems justifiable.

We compared the concatenated extant fragments of DNA REP1_DM to their respective ancestral sequences and estimated the frequencies of all 12 possible single-nucleotide substitutions using a maximum-likelihood approach to include multiple and back substitutions at the same site. Details on a more general version of this approach including neighbor-dependent substitution processes have been discussed by Arndt *et al.* (2003b). Once the 12 substitution frequencies ($\mu_{i\rightarrow j}$) are established, the stationary GC content (GC*) can be easily computed (Arndt *et al.* 2003a). We can also assume strand complementarity and estimate six substitution frequencies ($\mu_{C:G\rightarrow T:A}$); under this scenario $\mu_{C\rightarrow T} = \mu_{G\rightarrow A}$, for instance. We use the 12 rate estimates of substitution frequencies for all of our analyses except our mathematical model of codon bias evolution, in which we assume strand complementarity and accordingly use the 6 rate estimates.

Since we have only a finite amount of sequence data from which to estimate the substitution frequencies, these estimates will be affected by statistical errors; we can estimate these errors by bootstrapping our data set. For a category with a total of *N* aligned base pairs we resample the data, drawing randomly and with replacement *N* pairs of aligned ancestral and daughter nucleotides. From this resampled sequence data we estimated the substitution frequencies and the GC content as above. We repeat this resampling procedure *M* times and from the *M* estimates of the above quantities calculate their standard deviation, which gives the statistical error due to the limited amount of sequence data. In our case, we found that $M = 500$ samples is sufficient to estimate those errors.

Insertion and deletion analysis: For each BLAST hit, we concatenated all deleted bases (present in the ancestral but missing in the daughter fragment) as well as all inserted bases (present in the daughter but missing in the ancestral fragment). If the combined length of the insertions or deletions for an individual hit was ≥10 bp, we calculated the GC content of the combined insertion and deletion (indel) event. To estimate a bias in the base composition of insertions or deletions, we compared the GC content of the inserted bases to the GC content of the aligned region in the daughter fragment and the GC content of the deleted bases to the GC content of the aligned region in the ancestral fragment.

Recombination rate estimation: A list of all 615 genes that had been localized in both the physical and genetic maps in Release 3 of the *D. melanogaster* genome was kindly provided by FlyBase (D. Sutherland, personal communication). Genes mapped to the X chromosome were excluded, and the recombination rates within heterochromatin and on the fourth chromosome were considered to be zero. Using only genes in the euchromatic portions of the autosomes, a third-order polynomial curve was fitted to the genetic distance as a function of physical distance for each chromosomal arm ($R^2 \geq 0.96$ for all arms), and recombination (centimorgans per megabase) was calculated as the derivative of this polynomial at a given nucleotide coordinate. Recombination rates estimates for any locus in the *D. melanogaster* genome are available at http://cgi.stanford.edu/~lipatov/recombination/recombination-rates.txt.

The recombination rate of each copy of DNAREP1_DM was

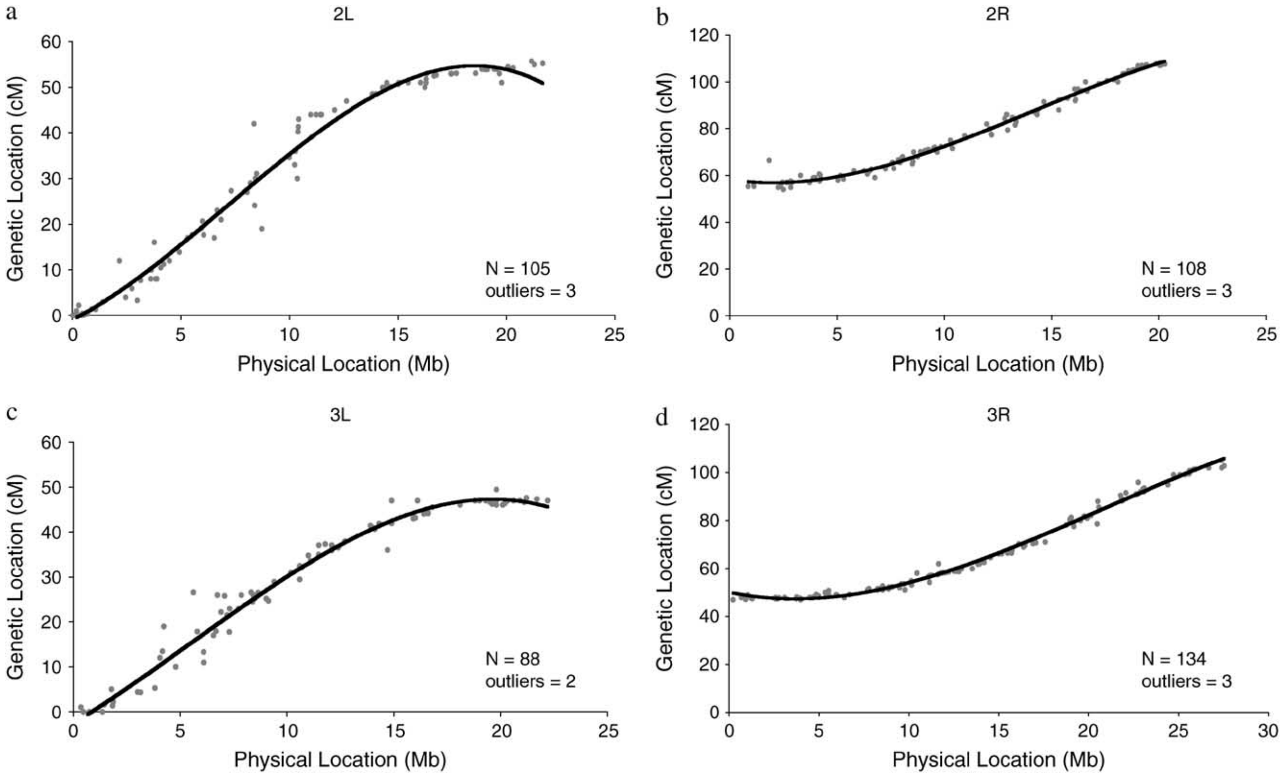

Figure 1.—Marey maps of chromosome arms (a) 2L, (b) 2R, (c) 3L, and (d) 3R of the *D. melanogaster* genome, based on the 615 genes in Release 3 of the genome sequence, which have been localized on both genetic and physical maps. A third-order polynomial curve mapping the genetic location onto the physical location is also represented. *N*, the numbers of genes from which the curve was derived; outliers, the number of genes excluded from the analysis.

estimated from the derivative of the polynomial curve using the midpoint of the element's sequence coordinates on the chromosome. Euchromatic copies were binned into two categories of recombination rate: those hits with recombination estimated as ≤2.3 cM/Mb were classified as "low recombination" and those hits where recombination was calculated to be strictly >2.3 cM/Mb were classified as "high recombination," following a previously reported recombination cutoff (Marais *et al.* 2001).

GC content of noncoding regions: For each gene on the autosomes of the *D. melanogaster* genome, we retrieved all intronic sequence, provided that the combined length of all introns for a particular gene exceeded 200 bp. In addition, we retrieved both 5′- and 3′-untranslated regions for all genes, provided that these regions exceeded 200 bp. We also retrieved 1000 bp "upstream" sequence for genes that were separated by >2 kb from their nearest neighbor 5′ of their transcription start site, as well as 1000 bp "downstream" sequence for genes that were separated by >2 kb from their nearest neighbor 3′ of the transcription termination site. Finally, we retrieved all "remaining" sequence that was >1 kb away from transcribed sites (and thereby not included in our upstream or downstream categories). GC-content of each fragment in our intron, 5′-UTR, 3′-UTR, upstream, downstream, and remaining categories was estimated excluding any N's in the sequence.

Coding sequences and codon usage: We retrieved coding sequences for all autosomal genes (n = 10,481) in Release 3.2 (FlyBase) of the *D. melanogaster* genome that were not located in telomeric regions (sections 21, 60–61, and 100) as defined by Bridges (1935). Genes mapped to heterochromatic contigs were not included in our analysis. For some genes several transcripts were listed, and for these genes we included only the first transcript listed in our data set, and both the protein length and codon bias estimates are based solely on this first listed transcript. For each autosomal gene we calculated overall optimal codon frequencies as defined by Duret and Mouchiroud (1999) in addition to calculating individual codon frequencies for genes >200 amino acids in length (n = 8308). Finally, for 9 amino acids we calculated the mean and standard error of individual codon frequencies of genes in both low- and high-recombination areas and used these distributions of codon frequencies in the following model as appropriate to test for selection.

Model of codon bias evolution: We developed a simple model of codon bias evolution that can be applied to amino acids encoded by only two codons, for which one codon is optimal and the other is nonoptimal. There are nine such amino acids: lysine, asparagine, aspartic acid, glutamine, glutamic acid, histidine, tyrosine, cysteine, and phenylalanine. In all cases, optimal codons end in G or C while nonoptimal codons end in A or U, respectively. For the purpose of this explanation, we examine the case where the optimal codon ends in G and the nonoptimal codon ends in A. The most basic substitution model is G:C $\xrightarrow{\mu_{\mathrm{nopt}} \times \rho_{\mathrm{nopt}}}$ A:T and A:T $\xrightarrow{\mu_{\mathrm{opt}} \times \rho_{\mathrm{opt}}}$ G:C, where μ_{nopt} is the substitution rate to a nonoptimal codon and ρ_{nopt} is the probability of fixation of such a substitution, while μ_{opt} and ρ_{opt} are the substitution rate to the optimal codon and the probability of fixation of such a substitution, respectively. At equilibrium, we expect $G \times \mu_{\mathrm{nopt}} \times \rho_{\mathrm{nopt}} = A \times \mu_{\mathrm{opt}} \times \rho_{\mathrm{opt}}$. If we define $\omega = \rho_{\mathrm{nopt}}/\rho_{\mathrm{opt}}$, then $\omega = (A \times \mu_{\mathrm{opt}})/(G \times \mu_{\mathrm{nopt}})$. Since we have distributions of A and G (from the 8308 autosomal genes), and μ_{opt} and μ_{nopt} (from our substitutional rate data assuming strand complementarity), we can draw from

TABLE 1

DNAREP1_DM distribution

Region	No. hits	Length hits (bp)
5′-UTR low	85	12407
5′-UTR high	4	574
3′-UTR low	26	4604
3′-UTR high	3	399
Intron low	328	51386
Intron high	13	2053
Upstream low	91	13941
Upstream high	11	1672
Downstream low	73	11485
Downstream high	4	620
Remaining low	399	63216
Remaining high	27	4079
Heterochromatin	392	63786

Distribution of DNAREP1_DM fragments on the autosomes of the *D. melanogaster* genome is shown. "Low" and "High" refer to recombination rates; see MATERIALS AND METHODS for "Region" definitions. For each region of the genome, the total number of DNAREP1_DM fragments is shown along with the total length of all hits in that region.

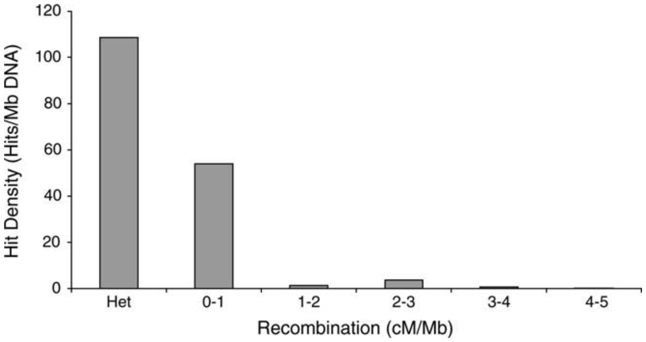

FIGURE 2.—The density of hits per megabase of DNA plotted against recombination estimates in centimorgans per megabase for the autosomes. Bins with numeric estimates of recombination are euchromatic sequence. Het., heterochromatic sequence.

these distributions to obtain a distribution for ω. We sampled from our distributions 1000 times to calculate the mean and standard deviations on our estimates of ω, and because we used the standard error in lieu of the standard deviation in generating distributions of codon frequencies and substitution rates, the standard deviation on our estimates of ω is actually equivalent to the standard error of our estimate of the mean. We can then estimate ω for low- and high-recombination areas and test whether this ratio of our fixation probabilities (ω) is significantly different in low *vs.* high areas of recombination. Under a background substitution model, the ratio of fixation probabilities should remain unchanged between recombination categories. We can then use our estimates of ω to infer $4N_es$ using Kimura's (1962) formula for the probability of fixation of a new mutation assuming codominance.

RESULTS

Recombination rates: We estimated recombination as a continuous variable for chromosome arms 2L, 2R, 3L, and 3R. We used all genes (615) that are localized on both genetic and physical maps (Release 3) of the *D. melanogaster* genome and fit the genetic position (centimorgans) to the physical location (megabases) using a third-order polynomial curve (Figure 1). Genes localized to the X chromosome, fourth chromosome, or heterochromatic sequence were not included in our analysis. We identified three, three, two, and three outliers on chromosome arms 2L, 2R, 3L, and 3R, respectively, by visual inspection and removed these genes from our analysis. For chromosome arms 2L, 2R, 3L, and 3R there were 105, 108, 88, and 134 genes that were used to generate the polynomial curve, respectively, and the polynomial curve fit these genes quite well in all cases: $R^2 \geq 0.96$ for all chromosome arms. Recombination was estimated as the derivative of the polynomial curve at a given nucleotide coordinate, following the example set by others (Kliman and Hey 1993; Comeron *et al.* 1999; Comeron and Kreitman 2000; Marais *et al.* 2001, 2003; Hey and Kliman 2002).

Distribution and abundance of DNAREP1_DM: DNAREP1_DM is the most abundant transposable element in the *D. melanogaster* genome, with copy number orders of magnitude higher than that of any other transposable element (Kapitonov and Jurka 2003). The consensus sequence for this element is 594 bp, and DNAREP1_DM is believed to be a nonautonomous derivative of a *Penelope*-like element due to the high (>70%) homology of a region of DNAREP1_DM to a 3′-UTR of a *Penelope*-like retrotransposable element (Kapitonov and Jurka 2003). DNAREP1_DM appears to have undergone a major burst of transposition several million years ago and has remained inactive ever since (Kapitonov and Jurka 1999, 2003; Singh and Petrov 2004).

Using BLAST parameters outlined in MATERIALS AND METHODS, we retrieved 1456 fragments of DNAREP1_DM from the autosomes of the *D. melanogaster* genome, which correspond to >230 kb of sequence (Table 1). These fragments are not evenly distributed across the genome and are preferentially located in regions of low recombination (Figure 2). This element was found on all chromosomes, and once we partitioned the hits on the basis of their proximity to genes, it became clear that not only were some hits in putative upstream or downstream regulatory regions (1000 bp 5′ or 3′ of a transcribed region) but also some hits were found within noncoding sequences of genes (Table 1). Copies inserted into introns were significantly more likely to fall on the antisense strand than on the sense strand ($P = 0.001$, *G*-test), while elements in 5′- or 3′-untranslated regions did not show this bias. Like elements in UTRs, elements inserted in upstream and downstream regions of autosomal genes showed no strand asymmetry relative to the direction of transcription of neighboring genes.

Background substitutional patterns of noncoding sequences: Our previous results strongly suggested that DNAREP1_DM went through a single major burst of transposition ~5 MYA and has remained inactive ever since (Singh and Petrov 2004). We therefore assumed a star-like phylogeny of DNAREP1_DM elements whereby each copy has been evolving independently since the time of its insertion. We assumed that the consensus sequence of all current DNAREP1_DM copies is a fair approximation of the ancestral DNAREP1_DM sequence. To infer background patterns of nucleotide substitutions in different types of noncoding DNA sequences on the autosomes of the *D. melanogaster* genome we compared current sequences of DNAREP1_DM with the presumed ancestral sequence of DNAREP1_DM. To correct for multiple hits we employed a maximum-likelihood approach (Arndt *et al.* 2003b). We estimated rates for each of 12 nucleotide substitutions in areas of high and low recombination for intron, UTR, upstream, and downstream regions of genes, as well as euchromatic sequence >1 kb away from a gene. We similarly estimated these substitutional rates for heterochromatic sequences. Additionally, as an internal control, we estimated rates of CpG (CG → CA and CG → TG) substitutions. Three low-complexity regions of the consensus sequence were filtered out of our analyses: bases 119–151 (33 bp), 242–279 (38 bp), and 482–504 (23 bp). However, among the remaining sections of the element, we did not find any substantial variability in base composition or substitution rate among different regions of DNAREP1_DM (data not shown).

We employed a hierarchical approach to investigate heterogeneity in background substitutional patterns across the autosomes of the *D. melanogaster* genome. First, we compared the substitutional profiles among elements inserted into introns, UTRs, upstream, downstream, and remaining regions within low- and high-recombination categories. We defined low recombination as ≤2.3 cM/Mb following a previously reported recombination rate cutoff (Marais *et al.* 2001). To test for rate heterogeneity among noncoding sequence types within these recombination regimes, we opted to use the predicted equilibrium GC content (GC*) as a proxy for the overall substitutional pattern to reduce the number of comparisons. For both low- and high-recombination areas, there were no significant differences in GC* among our noncoding DNA classes ($P \geq 0.09$, all pairwise comparisons, two-tailed *t*-test). As a result, within each recombination category, we pooled our elements into those falling into transcribed sequences and those falling into untranscribed sequences and tested for heterogeneity in substitutional profiles between these two classes of noncoding sequence. Because there was no significant difference in GC* between transcribed and untranscribed sequences ($P = 0.86$ and $P = 0.84$ for low- and high-recombination areas, respectively, two-tailed *t*-test), we pooled these categories as well. We then tested for variation in the substitutional spectrum among heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of euchromatin. GC* of heterochromatic regions (31.5%) was significantly lower than GC* of low-recombination areas of euchromatin (33.0%; $P = 0.006$, one-tailed *t*-test) and the GC* of low-recombination areas of euchromatin was nearly significantly lower than GC* of high-recombination areas (35.0%, $P = 0.055$, one-tailed *t*-test) (Figure 3).

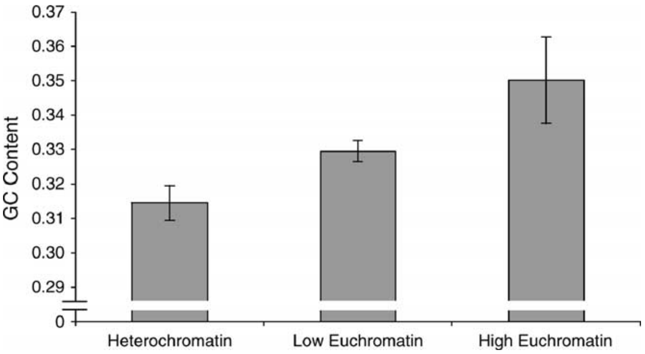

Figure 3.—Equilibrium GC content estimated from rates of nucleotide substitution for heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of euchromatin. Error bars correspond to the standard errors of our measurements.

To understand the underlying forces responsible for these differences in GC*, we more closely analyzed the differences in substitutional spectra among heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of euchromatin. Overall, the substitutional profiles for these three recombination regimes appear quite similar (Figure 4). While total substitution rate did increase with increased recombination rate, this increase was not statistically significant; total substitution rates for heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of recombination are 0.719, 0.725, and 0.737, respectively ($P = 0.43$ for heterochromatin *vs.* low-recombination euchromatin and low-recombination euchromatin *vs.* high-recombination euchromatin, one-tailed *t*-test). One nucleotide substitution's rate was significantly different in one comparison: the rate of A-to-G substitutions was significantly higher in low-recombination areas of euchromatin as compared to heterochromatin ($P \ll 0.05$, Bonferroni-corrected two-tailed *t*-test).

Base composition of indels: We compared the base composition of deleted and inserted bases within our autosomal DNAREP1_DM fragments to the base compositions of the aligned regions of the ancestral and daughter sequences, respectively, to assess potential base composition biases of indel events. While the average GC content of deleted bases (52.0%) is significantly higher than the average GC content of the alignable ancestral sequence (45.9%; $P \ll 0.0001$, paired two-tailed *t*-test), average GC content of inserted bases (35.0%) is significantly lower than average GC content of the alignable

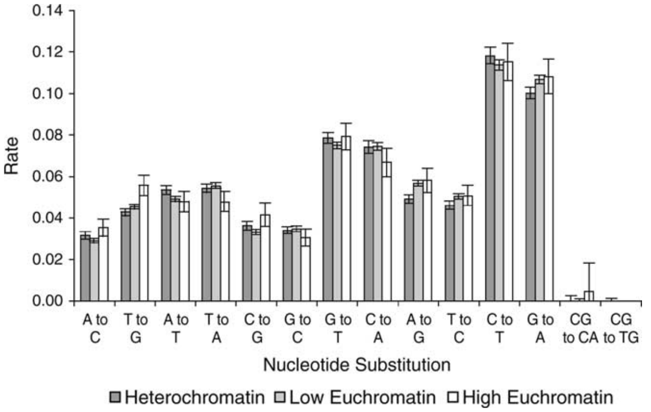

FIGURE 4.—Rates of nucleotide substitution inferred by comparing ancestral and daughter fragments of DNAREP1_DM for heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of euchromatin. Error bars denote standard error.

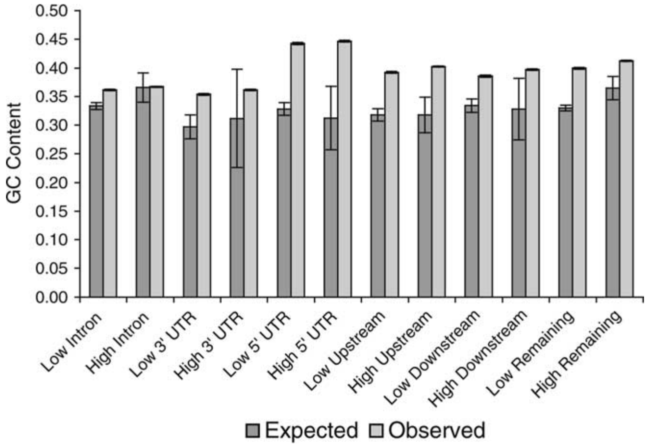

FIGURE 5.—Observed *vs.* expected GC content of noncoding regions. Expected GC content is estimated from nucleotide substitution rates; observed GC content was calculated from these noncoding regions directly. Error bars denote standard error.

daughter fragment (42.0%; $P = 0.036$, paired two-tailed *t*-test). It therefore appears that small indels in DNA REP1_DM have an overall bias toward reducing GC content of remaining sequences. However, there is no significant correlation between the GC content of either insertions or deletions and recombination rate (Kendall's $\tau = 0.078$ and 0.045; $P = 0.57$ and 0.43 for insertions and deletions, respectively).

Strand noncomplementarity: Using the substitutional profiles inferred from autosomal DNAREP1_DM fragments in heterochromatin, low-recombination areas of euchromatin, and high-recombination areas of euchromatin, we tested for asymmetry in rates of pairs of complementary nucleotide substitutions. In low-recombination areas, where the error on our estimates of substitution rates is the smallest, four of six pairs of complementary substitutions showed significant rate differences: A:T to C:G ($P \ll 0.0001$, two-tailed *t*-test), A:T to T:A ($P = 0.0007$, two-tailed *t*-test), A:T to G:C ($P = 0.0004$, two-tailed *t*-test), and C:G to T:A ($P = 0.01$, two-tailed *t*-test). The rate of T to G exceeded the rate of A to C, the rate of T to A exceeded the rate of A to T, the rate of A to G exceeded the rate of T to C, and the rate of C to T exceeded the rate of G to A (Figure 4). We were also able to detect significant differences in the same direction in the substitution rates of the A:T to C:G pair in heterochromatin and high-recombination areas of euchromatin ($P \ll 0.0001$, $P = 0.0002$, respectively, two-tailed *t*-test) and the C:G to T:A pair in heterochromatin ($P < 0.0001$, two-tailed *t*-test).

Base composition of noncoding DNA: To examine the effect of background substitutional patterns on the base composition of noncoding DNA, we compared the expected equilibrium GC content (GC*) of noncoding regions as inferred from our autosomal substitutional rate data to the observed GC content of those regions on the autosomes of the *D. melanogaster* genome (Figure 5). Expected GC* does increase with recombination for most noncoding sequence types, although not significantly ($P > 0.05$ all comparisons, one-tailed *t*-test). The observed GC content of all noncoding sequences in the genome does increase significantly with recombination ($P \leq 0.002$ all comparisons, one-tailed *t*-test). In addition, in most cases, the observed GC content is significantly higher than would be expected on the basis of our estimates of point substitution rates ($P \leq 0.02$, two-tailed *t*-test); the only cases where the increase is not significant are the intron, 3′-UTR, and downstream sequences in regions of high recombination.

Codon usage: Though not significant, we did detect differences in rates of background substitutions between low- and high-recombination areas of euchromatin on the autosomes of the *D. melanogaster* genome. In an effort to determine whether these substitutional differences were reflected in codon usage patterns, we examined codon usage patterns for all genes on the autosomes ($n = 10,481$ genes). On average, optimal codon frequencies of genes in areas of high recombination (53.9%) were significantly higher than optimal codon frequencies of genes in low areas of recombination (52.8%; $P < 0.0001$, one-tailed *t*-test) and using AKASHI's (1995) definition of major codons did not change these results. We also analyzed codon usage patterns of each gene >200 amino acids ($n = 8308$ genes) for all amino acids individually, and these data are presented in Table 2. Analysis of all genes (including genes <200 amino acids) did not qualitatively change our results. While the frequencies of many optimal codons do increase with recombination, many do not, and there do not appear to be general trends with respect to recombination for GC-ending or AU-ending nonoptimal codon frequencies.

Testing for selection: Our estimates of rates of background substitutions from areas of low and high recombination suggest that there may, in fact, be a substitu-

TABLE 2

Codon frequencies

Amino acid	Codon	Low recombination	High recombination	Amino acid	Codon	Low recombination	High recombination
Arg	*CGC*	*0.309*	*0.315*	Val	*GUC*	*0.233*	*0.232*
	CGG	0.157	0.149		*GUG*	*0.460*	*0.479*
	CGA	0.160	0.156		GUA	0.115	0.104
	CGU	*0.147*	*0.157*		GUU	0.193	0.184
	AGA	0.102	0.099				
	AGG	0.125	0.123	Lys	*AAG*	*0.686*	*0.699*
					AAA	0.314	0.301
Leu	*CUC*	*0.153*	*0.154*				
	CUG	*0.408*	*0.431*	Asn	*AAC*	*0.558*	*0.541*
	CUA	0.094	0.090		AAU	0.441	0.458
	CUU	0.109	0.095				
	UUA	0.054	0.046	Gln	*CAG*	*0.684*	*0.701*
	UUG	0.182	0.184		CAA	0.315	0.298
Ser	*UCC*	*0.247*	*0.252*	His	*CAC*	*0.604*	*0.597*
	UCG	*0.186*	*0.198*		CAU	0.391	0.400
	UCA	0.094	0.088				
	UCU	0.091	0.079	Glu	*GAG*	*0.656*	*0.671*
	AGU	0.141	0.141		GAA	0.343	0.329
	AGC	0.241	0.242				
				Asp	*GAC*	*0.476*	*0.454*
Thr	*ACC*	*0.383*	*0.394*		GAU	0.524	0.545
	ACG	0.247	0.251				
	ACA	0.190	0.183	Tyr	*UAC*	*0.632*	*0.618*
	ACU	0.180	0.171		UAU	0.366	0.380
Pro	*CCC*	*0.343*	*0.355*	Cys	*UGC*	*0.674*	*0.695*
	CCG	0.280	0.283		UGU	0.300	0.278
	CCA	0.238	0.239				
	CCU	0.137	0.123	Phe	*UUC*	*0.603*	*0.625*
					UUU	0.397	0.374
Ala	*GCC*	*0.441*	*0.457*				
	GCG	0.183	0.183	Ile	*AUC*	*0.448*	*0.472*
	GCA	0.174	0.168		AUA	0.204	0.194
	GCU	0.202	0.192		AUU	0.338	0.334
Gly	*GGC*	*0.407*	*0.422*				
	GGG	0.091	0.076				
	GGA	0.297	0.290				
	GGU	0.205	0.213				

Codon frequencies for all amino acids with synonymous codons in areas of high and low recombination as calculated from autosomal genes. Codons in italics denote optimal codons as defined by Duret and Mouchiroud (1999).

tional bias associated with recombination toward G and C. To test whether the substitutional bias that we found was sufficient to explain the 1.1% increase in optimal codon frequency between low and high areas of recombination, we examined the fixation probabilities of nonoptimal and optimal codons in our two recombination regimes. If the detected bias in background substitutional patterns is wholly sufficient to explain codon usage patterns, the ratio of the probability of fixation of the nonoptimal codon to the probability of fixation of the optimal codon should be unchanged across recombination categories. We restricted the application of our model to those amino acids encoded by two codons because codon usage data for threefold, fourfold, and sixfold degenerate amino acids do not fit a simple model of one or two equivalent, preferred codons with the remaining codons equally unpreferred (Table 2) and we were reluctant to fit more complicated models with multiple fitness parameters. We estimated this ratio of fixation probabilities of nonoptimal *vs.* optimal codons (ω) for the nine two-state amino acids in areas of both high and low recombination and tested for differences in our estimates of ω between recombination areas. For all nine amino acids, there were no significant differ-

TABLE 3

Fixation probability ratios and selection coefficients of nine amino acids

	Low recombination		High recombination		
Amino acid	ω	$4N_es$	ω	$4N_es$	*P*-value (ω)
Asn	0.383 (0.009)	0.959	0.409 (0.034)	0.894	0.230
Asp	0.533 (0.012)	0.629	0.586 (0.049)	0.535	0.152
Cys	0.216 (0.005)	1.534	0.193 (0.017)	1.643	0.102
Gln	0.223 (0.005)	1.500	0.207 (0.018)	1.577	0.185
Glu	0.253 (0.006)	1.373	0.238 (0.021)	1.437	0.234
His	0.314 (0.008)	1.157	0.325 (0.028)	1.124	0.356
Lys	0.222 (0.005)	1.504	0.209 (0.018)	1.567	0.235
Phe	0.320 (0.007)	1.141	0.292 (0.025)	1.231	0.148
Tyr	0.281 (0.006)	1.270	0.298 (0.026)	1.209	0.254

Ratio of fixation probabilities (ω) of nonoptimal codons and optimal codons for nine amino acids is shown. Standard errors on the estimates of these ratios are in parentheses and the *P*-values (one-tailed *t*-test) of testing for differences between estimates of ω from areas of low recombination *vs.* high recombination are also shown. The strength of selection ($4N_es$) inferred from these fixation probabilities is given for each amino acid in both recombination categories.

ences in our estimates of ω, and the estimated selection coefficients were small ($4N_es$ varied from 0.535 to 1.643) in both low and high areas of recombination (Table 3). Indeed, our model suggests that to reject a background substitutional model, optimal codon frequencies must increase in excess of 4%, in contrast to the observed ~1–2% increases (Table 2). In addition, there appears to be no overall difference in the strength of selection between areas of high and low recombination with respect to the fixation of optimal *vs.* nonoptimal codons ($P = 0.99$, paired sign test).

DISCUSSION

Rates of recombination: Using Release 3.2 of the *D. melanogaster* genome from FlyBase, we estimated recombination as a continuous function across a chromosome arm. Using a third-order polynomial curve, we fit the physical location (in base pairs) to the genetic position (in centimorgans) using a set of genes that had been localized on both maps. For all chromosome arms, there were very few outliers (overall <2%) and the fit of the polynomial curve to the rest of the data was quite good with $R^2 > 0.96$ for all chromosome arms. While there are obvious problems with estimating recombination as a continuous function, and while several other approaches have been employed in the past (Kindahl 1994; Carvalho and Clark 1999; Hey and Kliman 2002), the good fit we obtained gave us confidence that our recombination rates do in fact correspond to the relative rates of recombination in different regions of the genome. Our gradient of recombination values captures very low recombination (heterochromatin), low-recombination areas of euchromatin (≤2.3 cM/Mb), and high-recombination areas of euchromatin (strictly >2.3 cM/Mb).

Genomic DNAREP1_DM distribution: Because we are trying to quantify variation in background substitutional patterns associated with recombination such that it can be used as a null model against which a selective hypothesis for codon bias maintenance can be tested, and given the observation that the positive correlation between codon bias and recombination is unique to the autosomes of the *D. melanogaster* genome (N. D. Singh, J. C. Davis and D. A. Petrov, unpublished data), we restricted our analysis of background substitutional patterns to the autosomes of the *D. melanogaster* genome. On the basis of our BLAST parameters, we retrieved >1400 fragments of DNAREP1_DM from the autosomes of the *D. melanogaster* genome, corresponding to >230 kb of sequence. We partitioned the hits into those that fell either into heterochromatin or into euchromatin on the basis of the classifications in Release 3.2 of the *D. melanogaster* genome in which heterochromatic autosomal contigs are designated 2h and 3h. While some heterochromatic sequence may be included in contigs 2L, 2R, etc., the inclusion of this heterochromatic sequence within the euchromatic sequence category makes our comparisons between heterochromatin and euchromatin conservative.

The observation that DNAREP1_DM mobilized several million years ago (Kapitonov and Jurka 2003; Singh and Petrov 2004) coupled with the knowledge that noncoding DNA in *D. melanogaster* is lost quite rapidly (Petrov *et al.* 1996; Petrov 2002) suggests that most DNAREP1_DM copies will be riddled with internal deletions. This makes identifying full-length elements almost impossible given our BLAST criteria. Although we did merge overlapping hits, multiple fragments extracted from our BLAST alignments may in fact correspond to different segments of the same copy. To circumvent this issue of nonindependence, we concatenated all fragments within recombination categories to estimate rates of nucleotide substitutions. As a result, the only metric that may be affected by having individual copies retrieved as multiple fragments is the hit density (the number of hits per megabase; Figure 2). Because the rate of DNA loss is similar between heterochromatic and euchromatic regions (Blumenstiel *et al.* 2002), the result may simply be a systematic overestimation of the number of hits per megabase in all recombination categories, which does not qualitatively affect our results. Independent of the absolute values of hit density, there is a dramatic trend toward decreased hit density with increased recombination (Figure 2). DNAREP1_DM fragment density is highest in heterochromatic regions and declines rapidly in euchromatic sequence with increasing recombination. Given that large insertions may be mildly deleterious (Nuzhdin 1999; Petrov *et al.* 2003), we do expect to see a decrease in the number of copies that reach fixation with increased recombination due to Hill-Robertson effects and a probable increase in the density of functionally important sites in the regions of higher recombination. In addition, this pattern is consistent with the higher density of other transposable elements in the regions of low recombination in Drosophila (Bartolome *et al.* 2002; Blumenstiel *et al.* 2002).

DNAREP1_DM fragments are found in all types of noncoding regions, both genic and nongenic (Table 1). Fragments that do fall within introns of autosomal genes are significantly more likely to fall on the antisense strand than on the sense strand. This agrees well with previous observations and the accompanying hypothesis that regulatory motifs of retrotransposable elements are less likely to interfere with proper gene regulation if they are coded in the opposite direction (Smit 1999; Medstrand *et al.* 2002).

Background substitutional patterns: We used a maximum-likelihood procedure (Arndt *et al.* 2003b) to estimate rates for each of the 12 nucleotide substitutions, as well as CpG substitution rates from the comparison of the ancestor and daughter DNAREP1_DM sequences. We assumed a star-like phylogeny and used the consensus DNAREP1_DM sequence as the presumed ancestral sequence for each extant copy. This assumption of a star phylogeny has been thoroughly tested previously and the distribution of pairwise distances between extant copies DNAREP1_DM and the presumed ancestral sequence do support a single burst of transposition (Singh and Petrov 2004). DNAREP1_DM fragments are on average 15.2% diverged from the ancestral sequence, which suggests that the burst of transposition occurred ~4.6 MYA (Singh and Petrov 2004). Given this level of divergence, we expect that only 2.3% of sites have undergone multiple substitutions, which suggests that our inferences regarding the evolutionary history of these elements have not been compromised by parallel changes at individual sites. Even if the true phylogeny of these elements was not star-like, any departure from this model with respect to transposition would likely affect elements scattered across the genome. Given that we are specifically comparing substitutional patterns among elements in different regions of recombination, we do not expect a non-star-like phylogeny to systematically bias our results.

One potential concern with our methodology is that substitutional patterns inferred from DNAREP1_DM may not be reflective of substitutional patterns across the remainder of the genome. In particular, rates of gene conversion for multiple-copy elements such as DNAREP1_DM may be elevated relative to single-copy sequences. However, this will primarily affect rates of gene conversion when the elements are very young; as each copy evolves independently and accumulates substitutions the probability of ectopic recombination or gene conversion will decrease. Thus, while multiple-copy elements may evolve differently from single-copy sequences at their inception, this difference should be transient. Patterns of both nucleotide substitutions and indel dynamics deduced from single-copy nuclear transpositions of mitochondrial sequences are similar to those found in multiple-copy sequences such as transposable elements and pseudogenes in *D. melanogaster* (Petrov 2002; Singh and Petrov 2004). In addition, population genetic analysis of a fourth chromosome locus in *D. melanogaster* containing a DNAREP1_DM copy revealed no differences in mutational patterns between the DNAREP1_DM fragment and its noncoding flanking sequences (Singh and Petrov 2004). As a result, it seems likely that the substitutional patterns as estimated from DNAREP1_DM fragments do largely reflect substitutional processes across the genome.

We used equilibrium GC content (GC*) as a metric that reflected the relative strengths of individual substitutional forces to test for heterogeneity in substitutional patterns. Background rates of nucleotide substitutions were not significantly different across types of noncoding sequence; that is, the substitutional profile inferred from elements contained in introns was comparable to the substitutional profile estimated from DNAREP1_DM fragments far from genes, for example. This uniformity of the substitutional spectrum is consistent with our understanding of substitutional patterns in Drosophila, as substitutional patterns estimated from other noncoding sequences have also proven to be indistinguishable from one another (Singh and Petrov 2004). Because our estimates of substitutional rates were internally consistent, we pooled autosomal hits from different noncoding regions together for each recombination category.

We did detect significant differences in the background substitutional spectrum as approximated by GC* in the three regions of differing recombination rate. GC* in heterochromatic regions was significantly lower than GC* in low-recombination areas of euchromatin, which in turn was nearly significantly lower than the GC* in high-recombination areas of euchromatin (Figure 3). Therefore, it does appear that there is a substitutional bias toward G and C associated with recombination on the autosomes. This idea has been suggested previously from codon usage patterns and noncoding GC content (Marais *et al.* 2001), but our estimates of GC* are inferred directly from rates of nucleotide substitutions in noncoding DNA.

When we examined the rates of each of the 12 nucleotide substitutions more closely, we found few significant differences (Figure 4). Total substitution rate did not significantly increase with recombination, and only one substitution's rate increased with recombination: the rate of A-to-G substitutions was higher in low-recombination areas of euchromatin as compared to heterochromatin. This suggests that while the rates of individual nucleotide substitutions may not necessarily change significantly with recombination, the relative strengths of these individual nucleotide substitutions can in fact lead to significant differences in equilibrium GC content across recombination regimes.

The lack of significant increase of the substitution rates with recombination is consistent with previous results showing similar rates of synonymous divergence in Drosophila genes located in genomic regions with varying recombination rate (Begun and Aquadro 1992; Betancourt and Presgraves 2002). This contrasts with the pattern seen in primates, in which ∼6% of substitutions are apparently associated with the median recombination rate of 1.2 cM/Mb (Hellmann *et al.* 2003). However, given the errors on our estimates of total substitution rates in heterochromatin and low- and high-recombination areas of euchromatin, we are unable to reject the hypothesis that total substitution rates in *D. melanogaster* increase $< \sim 8\%$ between recombination categories. As a result, while we did not detect significant associations between recombination rate and nucleotide substitution rate, it is not clear whether recombination is associated with higher nucleotide substitution rates in *D. melanogaster.*

Our estimates of nucleotide substitution rates also corroborated previous reports of substitutional patterns in Drosophila. As was found in other noncoding sequences of *D. melanogaster* (Petrov and Hartl 1999), the C:G → T:A transition is the most prevalent substitution, and overall transitions occur more frequently than transversions, regardless of the location and recombination regime of the noncoding sequence (Figure 4). However, note that the T:A → C:G transition occurs at a similar rate to transversions. In addition, with respect to CpG substitutions, estimates for CG → TG and CG → CA substitutions are indistinguishable from zero (Figure 4). This is consistent with the expectation that these rates should be close to zero, as there is no known methylation in the *D. melanogaster* germline (Lyko *et al.* 2000). Furthermore, we found no significant differences in the rates of nucleotide substitutions between transcribed and untranscribed sequences, which is expected given the lack of transcription-coupled repair in Drosophila (De Cock *et al.* 1992; Van Der Helm *et al.* 1997; Sekelsky *et al.* 2000).

We were able to detect noncomplementarity in the rates of four of the six pairs of complementary nucleotide substitutions. Rates of T to G, T to A, A to G, and C to T significantly exceeded the rates of their complementary substitutions. It is not immediately clear why this is the case, and it is possible that there are yet unknown neighbor-dependent effects for nucleotide substitutions in Drosophila or that there are strand biases with respect to origins of replication, either of which could lead to noncomplementarity.

Implications for base composition of noncoding DNA: Although we have documented a recombination-associated substitutional bias toward G and C on the autosomes of the *D. melanogaster* genome, the role this bias plays in modulating the base composition of noncoding DNA is not clear. Given the rapid loss of nonfunctional sequences from the genome (Petrov *et al.* 1996; Petrov 2002) and the resulting high rate of sequence turnover at unconstrained intergenic loci (Singh and Petrov 2004) there is little expectation that nonfunctional sequences in the genome should be at equilibrium with respect to single-nucleotide substitutions. Indeed, this seems to be the case, as in almost all instances the expected equilibrium GC content of noncoding sequences inferred from nucleotide substitution patterns is significantly lower than the observed GC content (Figure 5). The base composition of insertion and deletion events may also contribute to the GC content of noncoding sequences, and our analyses suggest that small indels appear to decrease GC content of intergenic sequences. While small deletions appear to preferentially remove GC-rich sequences, small insertions tend to add AT-rich sequences. It is possible that large indels, however, are biased toward increasing GC content. At least with respect to large insertions, the high GC content of coding sequences may result in the large insertions from pseudogene formation and transposable element activity adding GC-rich sequences into the genome. Furthermore, variation in recombination rates over evolutionary time may also contribute to the departure from equilibrium with respect to GC content. Indeed, crossover frequencies in related species of Drosophila appear to be higher than those found in *D. melanogaster* (True *et al.* 1996; Takano-Shimizu 1999), which may suggest a recent reduction of rates of recombination in *D. melanogaster.* Accordingly, while the increases in expected GC content of noncoding sequences with recombination generally parallel increases in the observed GC content of those sequences (Figure 5), it is not clear that the relationship is causative.

Implications for codon usage: It has been extensively documented that codon usage bias is positively correlated with recombination rate in *D. melanogaster*, although this result now appears to hold only for the autosomes (N. D. Singh, J. C. Davis and D. A. Petrov, unpublished results). The absence of this correlation on the X chromosome led us to limit our quantification of background substitutional patterns to the autosomes, such that it would be an appropriate null model. While this positive correlation between recombination rate and codon bias was initially attributed to Hill-Robertson effects (Kliman and Hey 1993; Comeron *et al.* 1999; Hey and Kliman 2002), an alternative hypothesis was recently proposed (Marais *et al.* 2001, 2003). Given that 21 out of 22 optimal codons end in G or C, a substitutional bias toward G and C that is associated with recombination could putatively explain the increase in the frequencies of optimal codons with recombination. Although we have evidence in support of such a substitutional bias, we asked whether the magnitude of this bias was sufficient to explain the positive correlation between codon bias and recombination rate found in autosomal genes.

To test this hypothesis, we examined codon usage patterns for genes on the autosomes of the *D. melanogaster* genome. Interspecific sequence comparisons between *D. melanogaster* and *D. simulans* suggest that codon usage in *D. melanogaster* may not be in equilibrium, as there has been a strong reduction in codon bias in *D. melanogaster* that cannot be explained by mutational patterns (Akashi 1996; DuMont *et al.* 2004). Given that coding sequences are subject to repeated sampling of the nucleotide substitution process over a long period of time, however, any sequence in the genome that is close to equilibrium with respect to GC content will likely be coding. As a result, while the assumption of equilibrium of base composition may not apply for all coding sequences in *D. melanogaster*, we cannot even begin to address whether the magnitude of a substitutional bias toward increased GC is sufficient to explain recombination-associated patterns of codon usage in this species without the assumption of equilibrium.

As have others before us, we found that overall, optimal codon frequencies did increase with recombination. However, our estimate of a 1.1% increase in optimal codon frequency with increased recombination is lower than previously reported (Marais *et al.* 2001, 2003), which is due in large part to our exclusion of X-linked genes that have both higher recombination rates and higher codon bias (N. D. Singh, J. C. Davis and D. A. Petrov, unpublished data). In addition, when we investigated the response of individual codons to increased recombination, we found that while the frequencies of many optimal codons do increase with recombination, many optimal codons actually decrease in frequency with increased recombination (Table 2).

This inconsistency in the response of codon usage to increased recombination is difficult to reconcile with either selective or background substitutional models. If the positive correlation between codon bias and recombination rate were a result of selection on translational efficiency, then all optimal codons should increase in frequency with increased recombination, while all nonoptimal codons would decrease with increased recombination. A substitutional model would predict that all GC-ending codons would increase with increased recombination while AT-ending codons should decrease. Contrary to the predictions of both models, for several amino acids including asparagine, histidine, aspartic acid, and tyrosine, the GC-ending optimal codon decreases in frequency with increased recombination while the AT-ending nonoptimal codon concomitantly increases. For several additional amino acids, such as leucine, serine, threonine, proline, alanine, and glycine, while some GC-ending nonoptimal codons do decrease with recombination, others increase or do not change in frequency. Likewise, although certain AT-ending nonoptimal codons decrease with increased recombination, others increase or show no change in response to increased recombination.

Because of the diversity of responses of codon usage to recombination rate, it is likely that a number of additional factors also contribute to codon usage patterns in Drosophila. However, there is an overall positive correlation between recombination rate and level of codon bias on the autosomes of the *D. melanogaster* genome, and with the exception of five amino acids, all GC-ending optimal codons increase with recombination. We decided to inquire whether one needs to invoke Hill-Robertson effects above and beyond the differences in background substitutional patterns between regions of high and low recombination to explain recombination-associated patterns of codon usage. To address this question, we used individual codon frequencies for genes in low and high areas of recombination in combination with our substitutional rate data from DNAREP1_DM fragments and asked whether a background substitution model would be sufficient to explain the variation in codon usage associated with recombination.

Under a background substitution model, since selection is not differentially fixing nonoptimal and optimal codons in different areas of recombination, the ratio of the fixation probabilities of nonoptimal and optimal codons should remain unchanged between low- and high-recombination areas. If instead the efficacy of selection is increased with increased recombination, then the probability of fixation of an optimal codon will be increased and the probability of fixation of a nonoptimal codon should be decreased in highly recombining areas. Under this hypothesis, the ratio of the probability of fixation of the nonoptimal codon to the fixation probability of the optimal codon should be lower in regions of high recombination.

While the assumption of strand complementarity with respect to nucleotide substitutions was implicit in the construction of our mathematical model, we recognize the possibility that this assumption is not entirely appropriate. Although we were able to detect strand noncomplementarity in the rates of pairs of complementary nucleotide substitutions, we do not know how this asymmetry applies to coding regions of genes. Fortunately for this application, the differences in rates between complementary substitutions, albeit highly statistically significant, are quite small in magnitude (Figure 4). Therefore, although the assumption of strand complementarity is technically incorrect, our model provides good approximate estimates of fixation probability ratios and selection coefficients of optimal codons in *D. melanogaster*.

None of the nine amino acids to which we applied our mathematical model showed significantly different fixation probability ratios between low and high areas of recombination (Table 3). This suggests that at least for these amino acids, a background substitutional model is wholly sufficient to explain the relationship between codon bias and recombination rate on the autosomes of the *D. melanogaster* genome. It should be reiterated that while this substitutional bias may indeed reflect a mutational bias, the increase in the rate of substitutions toward G and C associated with recombination may also be mediated through biased gene conversion or selection on GC content of both functional and nonfunctional sequences across the genome.

Our inability to detect Hill-Robertson effects on codon usage in *D. melanogaster* does not necessarily imply that increased recombination rates do not increase the efficacy of selection on codon bias in this species. In fact, Hill-Robertson effects may be operating on a subset of genes, as suggested by Marais and Piganeau (2002). It is similarly possible that there is a recombination threshold above which linkage among sites is not sufficiently limiting to selection on codon bias (Kliman and Hey 2003). If this recombination threshold is below our recombination cutoff, then this may limit our ability to detect Hill-Robertson effects on codon usage. Furthermore, given that the time to equilibrium base composition may far exceed the time over which recombination rates have varied in this species, the base composition patterns we observe may reflect historical recombination rates as opposed to current ones, which could further explain our inability to detect Hill-Robertson interference.

Conclusions: We performed a genome-wide analysis of background rates and patterns of substitution on the autosomes of *D. melanogaster* using extant fragments of the transposable element DNAREP1_DM. This analysis allowed confirmation of the substitutional spectrum in this species and allowed us to detect subtle nuances within patterns of nucleotide substitution such as asymmetry in rates of pairs of complementary nucleotide substitutions. In addition, we found evidence in support of the nonequilibrium of noncoding sequences, which may reflect the dynamic turnover of noncoding sequences at unconstrained loci. Finally, we provide support for a recombination-associated substitutional bias toward G and C and examine the effects of this substitutional bias on other genomic properties, such as the base composition of noncoding and coding DNA. Fast sequence turnover due to the repeated processes of insertion and deletion in Drosophila (Singh and Petrov 2004) means that noncoding sequences are unlikely to be at equilibrium relative to the background pattern of nucleotide substitutions. It is thus difficult to assess the impact of the GC substitutional bias on the base composition of noncoding DNA, especially considering the AT-enriching bias in the base composition of small insertion and deletion events. However, given strong selection for genic function, we do not expect that coding sequences within the *D. melanogaster* genome would be subject to the same sequence turnover as unconstrained loci. As a result, the base composition of coding sequences may, in fact, be closer to equilibrium, and we can estimate how much of the recombination-associated variation in codon usage bias in *D. melanogaster* can be attributed to a recombination-associated substitutional bias toward G and C. Using codon frequencies for nine amino acids, we found that this background substitutional bias is entirely sufficient to explain differences in codon usage associated with recombination. However, these results do not discount the role of Hill-Robertson interference in modulating codon frequencies of other amino acids or even for particular classes of genes on the autosomes of the *D. melanogaster* genome. Finally, these results do not suggest that selection is not modulating codon frequencies in *D. melanogaster*. Rather, our data indicate that if indeed there is selection on codon usage, the efficacy of this selection does not noticeably increase with increased recombination.

The authors thank David Sutherland at FlyBase for providing a list of genes for use in our construction of Marey maps. Gabriel Marais and Jerel Davis proved invaluable throughout the duration of this project, and the authors gratefully acknowledge J. Michael Macpherson for his assistance with our model of codon usage. Richard Kliman, Mikhail Lipatov, and J. Michael Macpherson provided valuable feedback on our manuscript, and comments from two reviewers improved our manuscript considerably. This work was supported in part by the Stanford Genome Training Program (funded by 5 T32 HG00044 from the National Human Genome Research Institute).

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.

Aquadro, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. Dev. **7:** 835–840.

Arndt, P. F., C. B. Burge and T. Hwa, 2003a DNA sequence evolution with neighbor-dependent mutation. J. Comput. Biol. **10:** 313–322.

Arndt, P. F., D. A. Petrov and T. Hwa, 2003b Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol. Biol. Evol. **20:** 1887–1896.

Bartolome, C., X. Maside and B. Charlesworth, 2002 On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. Mol. Biol. Evol. **19:** 926–937.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in Drosophila melanogaster. Nature **356:** 519–520.

Betancourt, A. J., and D. C. Presgraves, 2002 Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA **99:** 13616–13620.

Blumenstiel, J. P., D. L. Hartl and E. R. Lozovsky, 2002 Patterns of insertion and deletion in contrasting chromatin domains. Mol. Biol. Evol. **19:** 2211–2225.

Bridges, C. B., 1935 Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. J. Hered. **26:** 60–64.

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

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

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

De Cock, J. G. R., E. C. Klink, W. Ferro, P. H. M. Lohman and J. C. J. Eeken, 1992 Neither enhanced removal of cyclobutane pyrimidine dimers nor strand-specific repair is found after transcription induction of the beta-3-tubulin gene in a Drosophila embryonic cell line Kc. Mutat. Res. **293:** 11–20.

DuMont, V. B., J. C. Fay, P. P. Calabrese and C. F. Aquadro, 2004 DNA variability and divergence at the Notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. Genetics **167:** 171–185.

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

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72:** 1527–1535.

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

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

Kapitonov, V. V., and J. Jurka, 1999 DNAREP1_DM. Repbase Update Release 3.4 (www.girinst.org/Repbase_Update.html).

Kapitonov, V. V., and J. Jurka, 2003 Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc. Natl. Acad. Sci. USA **100:** 6569–6574.

Kimura, M., 1962 On the probability of fixation of mutant genes in populations. Genetics **47:** 713–719.

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

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

Kliman, R. M., and J. Hey, 2003 Hill-Robertson interference in Drosophila melanogaster: reply to Marais, Mouchiroud and Duret. Genet. Res. **81:** 89–90.

Lyko, F., B. H. Ramashoye and R. Jaenisch, 2000 DNA methylation in Drosophila melanogaster. Nature **408:** 538–540.

Marais, G., and G. Piganeau, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes. Mol. Biol. Evol. **19:** 1399–1406.

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

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

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Medstrand, P., L. N. van de Lagemaat and D. L. Mager, 2002 Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. **12:** 1483–1495.

Nachman, M. W., 2002 Variation in recombination rate across the genome: evidence and implications. Curr. Opin. Genet. Dev. **12:** 657–663.

Nuzhdin, S. V., 1999 Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica **107:** 129–137.

Petrov, D. A., 2002 DNA loss and evolution of genome size in Drosophila. Genetica **115:** 81–91.

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

Petrov, D. A., E. R. Lozovskaya and D. L. Hartl, 1996 High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346–349.

Petrov, D. A., Y. T. Aminetzach, J. C. Davis, D. Bensasson and A. E. Hirsh, 2003 Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol. Biol. Evol. **20:** 880–892.

Sekelsky, J. J., M. H. Brodsky and K. C. Burtis, 2000 DNA repair in Drosophila: insights from the Drosophila genome sequence. J. Cell Biol. **150:** F31–F36.

Singh, N. D., and D. A. Petrov, 2004 Rapid sequence turnover at an intergenic locus in Drosophila. Mol. Biol. Evol. **21:** 670–680.

Smit, A. F. A., 1999 Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9:** 657–663.

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

True, J. R., J. M. Mercer and C. C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.

Van Der Helm, P. J. L., E. C. Klink, P. H. M. Lohman and J. C. J. Eeken, 1997 The repair of UV-induced cyclobutane pyrimidine dimers in the individual genes Gart, Notch and white from isolated brain tissue of Drosophila melanogaster. Mutat. Res. **383:** 113–124.

Communicating editor: N. Takahata