Patterns of Insertion and Deletion in Contrasting Chromatin Domains

Justin P. Blumenstiel, Daniel L. Hartl, and Elena R. Lozovsky

Department of Organismic and Evolutionary Biology, Harvard University

Transposable elements (TEs) play a fundamental role in the evolution of genomes. In Drosophila they are disproportionately represented in regions of low recombination, such as in heterochromatin. This pattern has been attributed to selection against repeated elements in regions of normal recombination, owing to either (1) the slightly deleterious position effects of TE insertions near or into genes, or (2) strong selection against chromosomal abnormalities arising from ectopic exchange between TE repeats. We have used defective non–long-terminal repeat (LTR) TEs that are ''dead-on-arrival'' (DOA) and unable to transpose in order to estimate spontaneous deletion rates in different constituents of chromatin. These elements have previously provided evidence for an extremely high rate of spontaneous deletion in Drosophila as compared with mammals, potentially explaining at least part of the differences in the genome sizes in these organisms. However, rates of deletion could be overestimated due to positive selection for a smaller likelihood of ectopic exchange. In this article, we show that rates of spontaneous deletion in DOA repeats are as high in heterochromatin and regions of euchromatin with low recombination as they are in regions of euchromatin with normal recombination. We have also examined the age distribution of five non-LTR families throughout the genome. We show that there is substantial variation in the historical pattern of transposition of these TEs. The overrepresentation of TEs in the heterochromatin is primarily due to their longer retention time in heterochromatin, as evidenced by the average time since insertion. Fragments inserted recently are much more evenly distributed in the genome. This contrast demonstrates that the accumulation of TEs in heterochromatin and in euchromatic regions of low recombination is not due to biased transposition but by greater probabilities of fixation in these regions relative to regions of normal recombination.

Introduction

For most of the last century, research in genetics tended to focus on the stability of genes from generation to generation and the stability of genomes through evolutionary time. More recently, rapid accumulation of data on genomic changes has shifted the emphasis from the stability of genomes to their potential plasticity. It is increasingly evident that transposable elements (TEs) can reshape genomes substantially and swiftly owing to their ubiquity among organisms and their inherent propensity for proliferation (e.g., SanMiguel et al. 1996). Differences in TE copy number can explain much of the variation in genome size even among closely related organisms (e.g., Kalendar et al. 2000). Transposable elements can be involved in generating large chromosomal rearrangements (Engels and Preston 1984; Lim and Simmons 1994; Caceres et al. 1999), can contribute to the evolution of protein-coding regions (Nekrutenko and Li 2001), and can even be recruited for novel functions in the host genome (Biessmann and Mason 1997; Smit 1999).

Heterochromatic DNA is enriched in TEs copies compared with euchromatic portions of the genome (Ananiev et al. 1984; Charlesworth, Lapid, and Canada 1992*a;* Charlesworth, Jarne, and Assimacopoulos 1994; Pimpinelli et al. 1995). However, the reasons for this asymmetrical distribution are not well understood. Several different scenarios can be entertained. For example, the deficiency of TEs in euchromatic regions with normal recombination might be a result of strong selection against chromosomal abnormalities arising from ectopic exchange between TE repeats (Montgomery, Charlesworth, and Langley 1987; Langley et al. 1988). The decreased abundance of TEs in regions of normal recombination also may be explained by the slightly deleterious position effects of TE insertions near or into genes (which are at higher density in euchromatin than in heterochromatin), coupled with an increased effectiveness of selection in these regions (Hoogland and Biemont 1996). There has been some difficulty in distinguishing between these hypotheses (Biemont et al. 1997; Charlesworth, Langley, and Sniegowski 1997). Yet another hypothesis for TE accumulation in heterochromatin involves the combined effects of specific targeting and positive selection for retention (Dimitri and Junakovic 1999). The three hypotheses are not necessarily mutually exclusive and may act together to cause the asymmetrical distribution of TEs in the genome.

In this study we explore the relationship between TE copy position, local rate of recombination, and rate and size of deletions. We examined non–long-terminal repeat (LTR) retrotransposons, which are ubiquitous among organisms, transpose through RNA intermediates, and usually create transpositionally inactive, ''dead-on-arrival'' (DOA) copies by way of transposition. These DOA copies provide an excellent model for studying the rates and patterns of spontaneous mutations, including insertions and deletions (indels) (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998; Lozovskaya et al. 1999). Large differences observed in the frequency and average size of spontaneous indels (indel spectra) in diverse organisms have led to the suggestion that spontaneous indels may have important implications for genome evolution with regard to the persistence of

Abbreviations: DOA, dead-on-arrival; indel, insertion and deletion; LTR, long-terminal repeat; TE, transposable element.

Key words: heterochromatin, mutation, transposable elements, genome size, *Drosophila melanogaster*, *Drosophila virilis*.

Address for correspondence and reprints: Daniel L. Hartl, Department of Organismic & Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138. E-mail: dhartl@oeb.harvard.edu.

Mol. Biol. Evol. 19(12):2211–2225. 2002

expendable sequences (reviewed in Hartl 2000). However, it is not known whether the indel spectra differ across classes of sequences (e.g., repetitive vs. unique) or chromatin domains (e.g., heterochromatin vs. euchromatin).

We estimated the deletion rates in euchromatin and heterochromatin in five different non-LTR retrotransposon families to avoid bias due to any particular features of individual elements. We demonstrated that rates of deletion are very high, and similar to earlier estimates, in regions of little or no recombination. This suggests that the high rates of deletion found in *Drosophila* are not biased by selection against ectopic exchange. Furthermore, we have shown that deletion rates are similar in regions of normal recombination and in heterochromatic regions with virtually no recombination. Based on these data, we derive an upper bound for $N_e s$ in the regions in which deletions were studied, and show that this limit is too small to bias the estimate of rates of deletion toward artificially high values.

We also performed an extensive analysis of age distributions of TE copies, both within TE families and within genomic regions. We found that there is great variation in the age distributions of different TE families. Furthermore, there is a tremendous difference in the age distributions of TE copies in different genomic regions. Very young (recently transposed) copies are found at similar densities across regions of different recombination rates. Older copies that transposed earlier, on the other hand, are substantially more prevalent in regions of low or no recombination. Based on these findings, we argue that non-LTR elements transpose nearly randomly into different parts of the genome. However, they fail to fix in regions of normal recombination, whereas in heterochromatin they age and accumulate with time. The relatively rapid clearance from regions with higher recombination is consistent with weak selection against DOA copies present in this portion of the genome (D. Petrov, personal communication). As for the mechanism of selection, the data do not distinguish between ectopic recombination and the slightly deleterious position effects. However, our data suggest that there might be some selection in favor of very large deletions in euchromatic regions with normal recombination.

The evolution of mobile DNA copy number is driven by new elements that invade the genome, propagate, and become inactivated by the host. Subjected to negative selection and elimination from euchromatin, they persist, grow older, deteriorate by mutation, and gradually accumulate in their genetic graveyard—the heterochromatin.

Materials and Methods

To determine the effect of genomic location on rates of deletions and TE retention, available *D. melanogaster* sequences from the Celera/Berkeley Drosophila Genome Project (BDGP) genome sequencing project were analyzed. For comparison, we also determined the sequences and genomic positions of a sample of TEs from *D. virilis*.

Retrieval of *D. melanogaster* TE Sequences

Transposable element sequences, along with 1,000 bp of available 5′ and 3′ flanking regions, were retrieved from the whole-genome database using BLAST with default settings from the National Center for Biotechnology Information (NCBI) web site (Celera Release October 2000 and available P1, BAC, and cosmid sequences from the BDGP). Sequences corresponding to five non-LTRs TEs were used in this study: *jockey* (Priimagi, Mizrokhi, and Ilyin 1988), *Waldo-A* (Busseau, Berezikov, and Bucheton 2001), *X* (Tudor et al. 2001), *Helena* (Petrov et al. 1995), and *You* (Berezikov and Busseau 2000).

To guard against potential inaccuracies in sequences of repeated elements in the *D. melanogaster* genome sequence, we performed a comparison of all available fragments and flanking sequences which could be identified in the final finished P1, BAC, and cosmid sequences from the BDGP with the October 2000 release of the whole-genome shotgun (WGS) sequence from Celera. Fragments from different data sets were matched by chromosomal positions and flanking sequences. Because the whole-genome sequence was not available in finished form from the BDGP project, this comparison was performed on a subset of fragments in our analysis ($n = 68$). We found near identity between flanking sequences from the two projects. Of the 68 flanking sequence sets of 2,000 bp each, there were no discrepancies between the two projects in 59 sequences, and there was a mean discrepancy between flanking sequences of 0.15 mismatches/kb (range: 0.0–1.0 mismatches/kb, counting each indel as a single mismatch). This comparison suggests that unique sequences are quite reliable from both projects.

However, we found that young fragments with less than 5% nucleotide difference between the active element and the genomic sequence were likely to contain many mismatches between the WGS and BDGP sequences. Of the 43 fragments identified in the WGS sequence that showed greater than 95% identity with the active copy, only seven showed no discrepancies between the two projects (mean fragment size compared, 2,629 bp; range 185–5,266 bp). The mean discrepancy between the two projects within these regions was 25-fold greater than that in flanking regions (3.79 mismatches/kb including each indel as one mismatch; range 0.0–14.2 mismatches/kb). Single base-pair substitutions and indels were the most common discrepancies between WGS and BDGP data sets. However, several larger deviations were also identified. In these 43 fragments, we identified eight indel discrepancies ranging from 2 to 10 bp and six indel discrepancies larger than 10 bp, including one of 49 bp, two of 51 bp, and one of 421 bp. Hence, young TE fragments, which have recently transposed and are therefore very similar to each other, appear to have been misassembled by the WGS software. Hereafter, we will call these very similar fragments ‘‘nearly exact’’ or NE fragments.

On the other hand, fragments that are quite different from the active element (greater than 5% sequence differences) matched much more closely between the two genome projects. Of the 25 fragments identified in the WGS sequence which showed less than 95% identity to the active copy, 20 showed no discrepancies between the two projects (mean fragment size, 589 bp; range: 67–2,631 bp). The mean discrepancy between the two projects within these regions was 0.32 mismatches/kb, counting each indel as one mismatch (range 0.0–3.14 mismatches/kb). Furthermore, of the only four indel discrepancies identified, all were a single base pair in size. Hence, it appears that older fragments with more sequence divergence were assembled in the two projects at a level of accuracy similar to that of the unique flanking sequences.

Because of the relatively high error rate found among NE fragments retrieved from the WGS sequence, these were not included in the rate of deletion-insertion analysis or were replaced (whenever possible) with the final finished sequence available from BDGP (http://www.fruitfly.org/sequence/assembly.html). Our phylogenetic analysis of all NE fragments available in the final finished sequence indicated less than 0.005 substitutions/bp. Because nearly all the NE fragments were greater than 99% identical to the active copy of the TE, in the analysis of the distribution of TE copies by age, these fragments were appropriately binned into the category with 0.0–0.02 substitutions/bp.

Genomic locations were available for TE fragments in euchromatin. Sequences that were not assigned to a euchromatic location by either genome-sequencing project were considered heterochromatic. Myers et al. (2000) pointed out that ‘‘a substantial fraction of these must be nonrepetitive islands within the heterochromatin of the centromeres, or may represent as-yet-unidentified foreign DNA.’’ We tested the possibility of foreign DNA and erroneous mapping experimentally: unique flanking sequences of five such unmapped fragments were used to generate PCR primers. Successful amplification of all five fragments indicated their presence in the *D. melanogaster* genome. These fragments were subsequently used for in situ hybridization with polytene chromosomes, and none were located in the euchromatic portion of the chromosome arms. Thus, we can confirm that most unmapped fragments from the genome sequence of *D. melanogaster* represent heterochromatic regions. A summary of fragment numbers and locations is given in table 1.

Generation of *D. virilis* TE Sequences

LambdaScan as well as P1 libraries were constructed and screened for the non-LTR element *Helena*. Positive clones were classified as either euchromatic or heterochromatic based on in situ hybridization of flanking regions with polytene chromosomes of a strain lacking *Helena* (Lozovskaya et al. 1999). After mapping, clones containing the *Helena* element were subcloned and sequenced.

Estimation of Indel Rates

Alignments were generated manually and also with the aid of the program BLAST2 (Tatusova and Madden 1999) on the NCBI web site. Genomic fragments were aligned in a pairwise fashion to the active element using the following settings: reward for match, 100; penalty for mismatch, 125; penalty for gap, 500; penalty for gap extension, 9. By making gap extension inexpensive relative to gap opening, fewer larger indels were favored over a large number of small indels. Furthermore, relative to the default settings, mismatch penalties were less expensive than match rewards. We encouraged a greater number of mismatches and fewer indels than the default settings to avoid overestimating the rate of indels. We found that these settings generated reasonable alignments, but all alignments were also examined by eye and adjusted as necessary.

Table 1
Number of Non-LTRs Analyzed

	Density Analysis				Indel Rate Analysis			
Element	H	E(LR)	E(NR)	All	H	E(LR)	E(NR)	All
Helena	16	7	7	30	4	5	2	11
jockey	7	15	62	84	6	8	2	16
You	9	9	2	20	6	5	1	12
Waldo-A . . .	23	41	12	76	18	22	4	44
X	18	23	13	54	9	13	1	23
Total.	73	95	168	264	43	53	10	106

Note.—Transposable element copies whose map position was not identified by either sequencing project were identified as heterochromatic (H). Low-recombination regions, denoted E(LR), were defined as those regions with rates of recombination less than 5×10^{-9} per base pair, corresponding to salivary polytene chromosome divisions 1A–2D, 20C–20F, 21A, 38A–44C, 60C–60F, 61A–61B, 75F–84F, and 101–102 (Kliman and Hey 1993; Fay and Wu 2000). Fragments residing in other regions were considered to be in regions of normal recombination, denoted E(NR). The numbers of fragments analyzed for density analysis was larger than those used for indel rate analysis. Recently transposed (‘‘young’’) fragments whose sequence was not confirmed with the finished sequence were not included in the indel rate analysis because young fragments were found to contain many discrepancies between the finished and the WGS sequences.

Aligned sequences were next subjected to phylogenetic analysis to determine the number of substitutions per base pair in each terminal branch. The rationale for using the number of terminal branch substitutions as a proxy for fragment age is outlined in Petrov, Lozovskaya, and Hartl (1996). In previous work, parsimony analysis was used because the TE fragments were more closely related than those in the present analysis. With more divergent copies, parsimony would tend to prefer homoplasies in the internal branches rather than parallel changes in the terminal branches. For this reason we used maximum likelihood (ML) analysis to estimate the number of terminal branch substitutions per base pair. The difference between parsimony and ML is quite small in any case. When parsimony analysis was compared with likelihood analysis, the latter nearly always gave slightly larger estimates of terminal branch substitutions per base pair, thus yielding somewhat more conservative estimates of indel rates. ML trees were created using PAUP* 4.0 b10 (Swofford 2002). Trees were searched heuristically using TBR branch swapping and optimized under an HKY85 + Gamma (with eight rates) model (Hasegawa, Kishino, and Yano 1985; Yang 1993). The number of terminal branch substitutions per base pair was estimated under the same model. Several other models of evolution were also examined: Jukes and Cantor (1969), Kimura's two-parameter (1980), and HKY85 without gamma. Estimates of the number of terminal branch substitutions per base pair were nearly identical under all these models.

Drosophila melanogaster TE fragments were classified according to their genomic location: heterochromatin, euchromatin with normal recombination, or euchromatin with low recombination. Low-recombination regions were defined as regions with rates of recombination less than 5×10^{-9} per base pair. These include salivary polytene chromosome divisions 1A–2D, 20C–20F, 21A, 38A–44C, 60C–60F, 61A–61B, 75F–84F, and 101–102. Rates of recombination and classification scheme were based on those used by several groups (Kliman and Hey 1993; Kindahl 1994; Charlesworth 1996*b*; Comeron, Kreitman, and Aguadé 1999; Fay and Wu 2000; Bartolomé, Maside, and Charlesworth 2002; Hey and Kliman 2002).

Rates of insertion and deletion were estimated assuming the Poisson process. The ML estimate of insertion is given by

$$\hat{\lambda} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} \alpha_i t_i} \qquad (1)$$

where n_i is the number of insertions in fragment *i,* α_i the average size of the fragment *i* ([initial size in bp + final size in bp]/2), and t_i the number of substitutions per base pair in fragment *i* as inferred from the ML assessment of terminal branch substitutions per base pair.

A slight correction must be made for estimating the rate of deletion due to the fact that for a given TE fragment inserted into the genome, there are a greater number of sites for observing small deletions than for observing large deletions. For example, in a 500-bp fragment there are 498 sites in which a 1-bp deletion can occur; however, there are only 298 sites for a 200-bp deletion (because a 200-bp deletion will not be observed if it overlaps with either end of the fragment). To make this correction, we estimated the rate for each deletion size and then summed over all rates. Thus, the ML estimator for deletion rate is

$$\hat{\lambda} = \sum_{j=1}^{MD} \frac{\sum_{i=1}^{m} n_{i,j}}{\sum_{i=1}^{m} \alpha_{i,j} t_i} \qquad (2)$$

where $n_{i,j}$ is the number of deletions of size *j* in fragment *i,* $\alpha_{i,j}$ the number of available sites for deletions of size *j* in fragment *i,* MD the maximum deletion size, and t_i the number of substitutions per base pair in fragment *i* as inferred from the ML assessment of terminal branch substitutions per base pair.

Ninety-five percent confidence intervals for rate estimates and mean deletion and insertion sizes were determined by 1,000 bootstrap replicates in which the data were sampled with replacement and used to re-estimate the parameter. Tests of significance for difference were performed by generating 1,000 bootstrap comparisons and determining the frequency of comparisons in which one parameter estimate was larger than the other parameter estimate.

Examination of TE Insertions for Polymorphism

We screened eight strains of *D. melanogaster* (Canton S, St. Louis, Oregon R, Cotonou, Harwich, Hikone R, Tokyo, Raleigh NC) for the presence or absence of 10 *jockey* insertions identified in the WGS sequence. Of these 10 inserts, five corresponded to young fragments located within the euchromatic region of normal recombination. Five corresponded to mature inserts, one residing in the euchromatic region of normal recombination and four residing in the heterochromatin.

Screening was performed by designing primers specific to sequences flanking the insert (determined from the WGS). Presence or absence of particular inserts was confirmed by PCR amplification.

Results

Distribution of TEs in *Drosophila* Genome

We found a greater TE abundance in regions of low or no recombination, as have many other studies (Charlesworth, Lapid, and Canada 1992*a,* 1992*b*; Rizzon et al. 2002). Figure 1 shows the density of TE copies within each genomic region, where density is defined as the number of TE fragments per Megabase. Problems with the assembly of repeat sequences makes a phylogenetic estimation of age impossible for all recently transposed sequences that are very similar to one another (see *Materials and Methods*). For this reason we classified as ''young'' those that were >98% identical to the active copy (mean fragment size, 1,709 bp; range 213–5,382 bp). Classification of such closely related sequences as ''young'' is supported by phylogenetic analysis of young copies whose sequence can be confirmed with the final finished BAC, P1, or cosmid sequence. Such analysis always confirmed an age of less than 0.005 substitutions/bp. All other fragments were classified as ''mature.''

Figure 1 shows that in heterochromatin, TE fragments are about 19 times denser than in euchromatic regions with normal recombination, and in euchromatic regions with low recombination, TEs are 4.5 times denser than in euchromatic regions with normal recombination. This skew in distribution is almost entirely due to the much higher density of older fragments in heterochromatin and regions of low recombination. Relative to regions of normal recombination, young fragments are only about 1.2-fold denser in regions of low recombination and only about 1.9-fold denser in heterochromatin. However, mature copies show a substantially greater skew. The density of mature TEs in heterochromatin is more than 90-fold greater than the density in regions of normal recombination, and the density of mature copies in regions of low recombination is about 18-

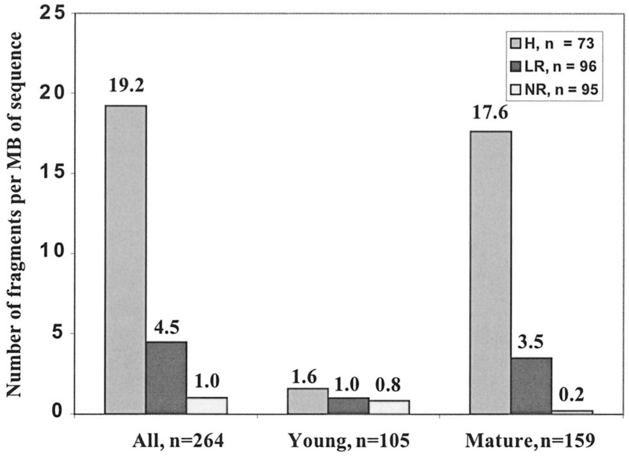


FIG. 1.—The density of TEs within different genomic regions in *D. melanogaster*. All fragments corresponding to the five non-LTR sequences were identified in complete genomic sequence. Density of fragments was determined based on the available sequence from that particular region and the number of fragments identified. NR (euchromatin, normal recombination), 94.9 Mb of sequence analyzed; LR (euchromatin, low recombination), 21.3 Mb of sequence analyzed; H (heterochromatin), 3.8 Mb of sequence analyzed.

fold greater than that in regions of normal recombination.

Age Distribution of TE Families

For non-LTR retrotransposons, Petrov, Lozovskaya, and Hartl (1996) showed that phylogenetic analysis could be used to distinguish between nucleotide changes that occur within an active TE lineage and those that occur within a fragment after transposition. Selective neutrality in substitutions that have occurred after insertion is inferred from the observation that the rate of substitution in terminal branches is equal among all codon positions. Similarly, purifying selection on coding sequences in the active lineages is inferred from an excess of third-position substitutions relative to the first and second positions in the internal branches. These patterns were observed when phylogenetic analysis was performed on the TE sequences in this study from both *D. melanogaster* and *D. virilis* (data not shown). Thus, the number of terminal branch substitutions can serve as a neutral benchmark for the age of a fragment. This allows the determination of rates of insertion and deletion as a function of neutral nucleotide substitutions.

Estimates of terminal branch substitutions per base pair within TE fragments can be used in turn to identify patterns of TE fixation within the genome through time. In figure 2 we have plotted the frequency of TE copies sorted by terminal branch substitutions per base pair. The total number of TEs analyzed is somewhat less than the number used in the density analysis because although WGS fragments which were 98% identical to the active copy could be classified reliably as falling in the 0.0- to 0.02-substitutions/bp range, the sequence of some fragments that were in the NE class could not be confirmed with the finished sequence, and these were not subjected to phylogenetic analysis.

There is a clear variation in the age distribution of fragments from different TE families. For example, more than 80% of *jockey* copies in the genome have less than 0.02 terminal branch substitutions/bp, and most of these reside in euchromatic regions of normal recombination. The rest of the *jockey* fragments are much older (0.07–0.14 terminal branch substitutions/bp), and these reside primarily in heterochromatin and regions of low recombination. The distribution of *You* copies tells a different story for its history of mobilization. The number of terminal branch substitutions per base pair for all *You* fragments is between 0.00 and 0.06. Furthermore, although older fragments primarily occupy regions of low recombination, there are fewer relative numbers of young fragments in regions of normal recombination.

Polymorphism of Fragment Insertions

We examined five mature inserts for evidence of polymorphism of insertion site within the genome of eight wild-type strains. All five mature copies were fixed in all the eight strains examined. Young copies, on the other hand, are polymorphic. One young insertion was found to be homozygous in two wild-type strains and absent in six others. A second young insertion was heterozygous in one strain and absent in the seven remaining strains. The three other inserts were absent in all eight strains. Thus, young inserts show a tendency to be polymorphic, and mature inserts, which are at much higher density in heterochromatin, show a tendency to be fixed.

Deletion Rates

There is a significant correlation between the number of substitutions per base pair and the number of deletions per base pair within TE fragments ($R^2 = 0.43$, $P < 1 \times 10^{-10}$, $n = 106$). Using equation (2), we calculated the rates of deletion in five different TE families (table 2). There is some variation in deletion rate among different TEs. For example, in *D. melanogaster* the deletion rate in *Helena* elements appears to be higher than in all other elements, and the 95% confidence intervals do not overlap. *Helena* fragments within *D. virilis* also show a greater rate of deletion, but the confidence interval overlaps with those of *X* and *You* in *D. melanogaster*. Deletion rates in the *You* element are also larger than rates within *jockey, Waldo-A* and *X,* although their 95% confidence intervals are largely overlapping. The rate of deletion is very similar among *jockey, Waldo-A,* and *X*. The apparently higher deletion rates found in *Helena* and *You* may result in part from the smaller sample size for these fragments ($n = 11$ and 12, respectively).

Because we were interested primarily in determining a genome-wide rate of deletion and potential variation among different regions of the genome, we combined all elements and estimated rates of deletion across the genome. This is reasonable because *Helena, jockey, Waldo-A, X,* and *You* are distributed in a similar manner among genomic regions. The slightly greater proportion

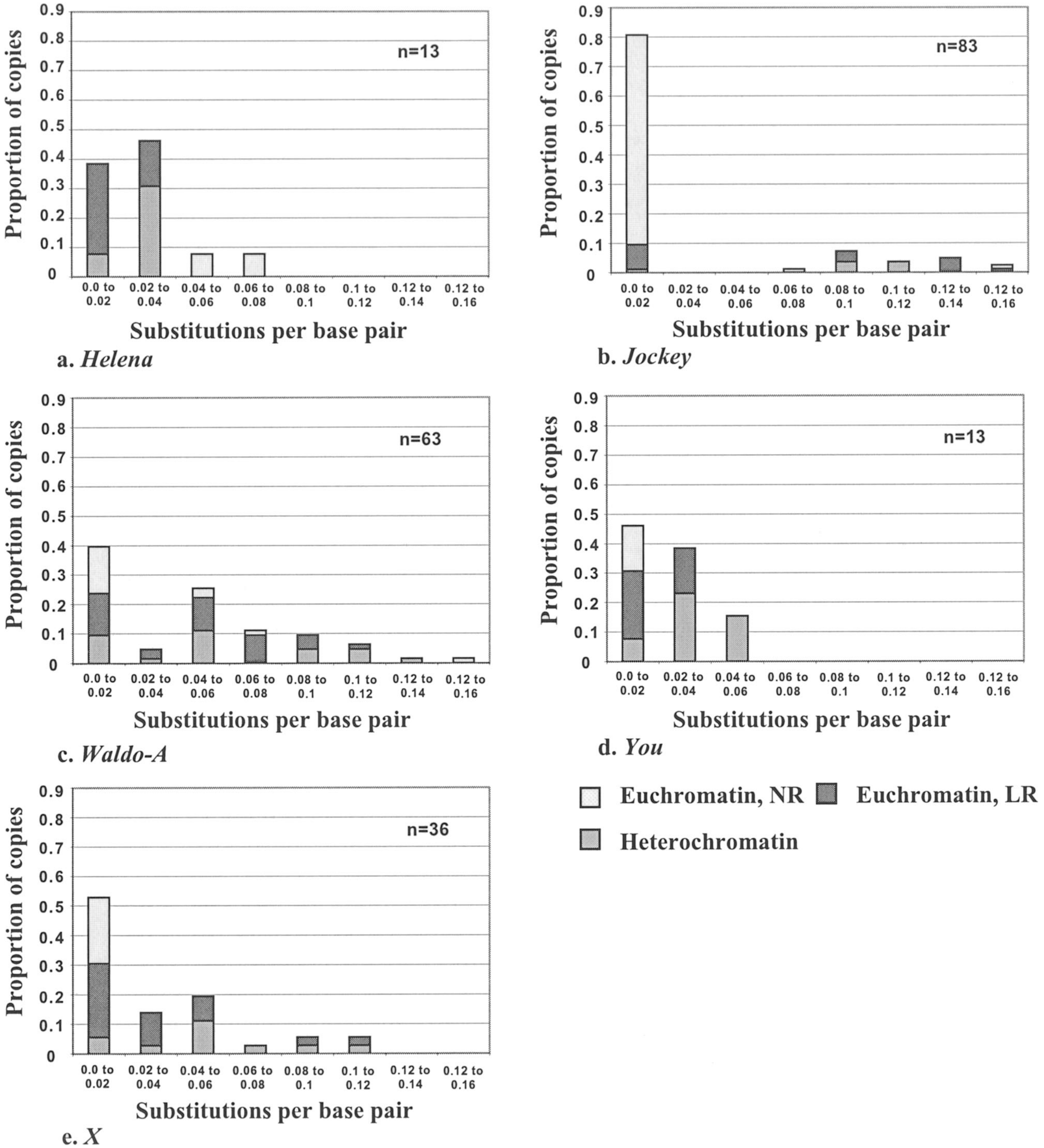


FIG. 2.—Age distribution of five different LTRs in the three regions of the genome of *D. melanogaster*. The number of terminal branch substitutions was determined by phylogenetic analysis. Fragments which were greater than 98% similar to the active copy were placed in the 0–0.02 substitution class because phylogenetic analysis performed on such fragments, whose sequence could be confirmed from the finished sequence, nearly always showed less than 0.005 terminal branch substitutions/bp from the active elements.

of *Helena* fragments in regions of normal recombination, relative to other regions, makes the test of selective neutrality of deletions conservative, owing to the greater rates of selectively driven clearance of TEs in regions of normal recombination.

Rates of deletion in *D. melanogaster* are very high in all three regions of the genome. They are also remarkably similar in heterochromatin, low recombination regions of euchromatin, and normal recombination regions of euchromatin (table 3). The 95% confidence intervals of the estimates of deletion rates are substantially overlapping. Because of the potential difficulties associated with estimating rates of recombination, we tested whether the similarity in rate estimates may have been

Table 2
Deletions per Substitution for Each TE Family with 95% Confidence Intervals

Species	Element	Deletion Rate (deletions per terminal branch substitution)
D. melanogaster	*Waldo-A*	0.099 (0.089 to 0.113)
	jockey	0.107 (0.094 to 0.128)
	X	0.110 (0.082 to 0.141)
	Helena	0.263 (0.200 to 0.381)
	You	0.153 (0.120 to 0.192)
D. virilis.	*Helena*	0.186 (0.136 to 0.248)

due to misclassification of cytological regions 19 and 20A–B as regions of normal recombination. When fragments residing in these regions were removed from the analysis, the remaining fragments were from regions 17A, 65B, 17D, 5E, and 15E, which are considered regions of high recombination according to Charlesworth (1996*b*) and Bartolomé, Maside, and Charlesworth (2002). The estimated deletion rate remained essentially the same (0.10 deletions per terminal branch substitution with 95% confidence interval of 0.083 to 0.124). Not only are deletion rates similar among different genomic regions, but we can also conclude with 99% confidence that rates of deletion in regions of normal recombination could not be more than 32% higher than those in heterochromatin. Even excluding the fragments residing in cytological sections 19 and 20A–B, we are able to reject a 32% higher rate of deletion with 99% confidence.

Rates of deletion are also high in the heterochromatic portion of the genome of *D. virilis*. Furthermore, the 95% confidence intervals for the deletion rates in euchromatin versus heterochromatin in *D. virilis* are overlapping. The rate of deletion in euchromatin is significantly greater ($P = 0.023$), but the sample size of euchromatic fragments is small ($n = 5$), so it is difficult to gauge whether a higher deletion rate in euchromatin is a true phenomenon. Furthermore, the apparently greater rate of deletion in euchromatin is due to one exceptional outlier fragment; when this is removed from the analysis, the difference is no longer significant ($P = 0.065$).

Distribution of Deletion Sizes

Figure 3 shows the distribution of deletion sizes by genomic region in *D. melanogaster* and *D. virilis*. The distributions are similar in all regions. The highest frequency of deletions is in the size range 1–10 bp (47.5% in *D. melanogaster* and 54.7% in *D. virilis*).

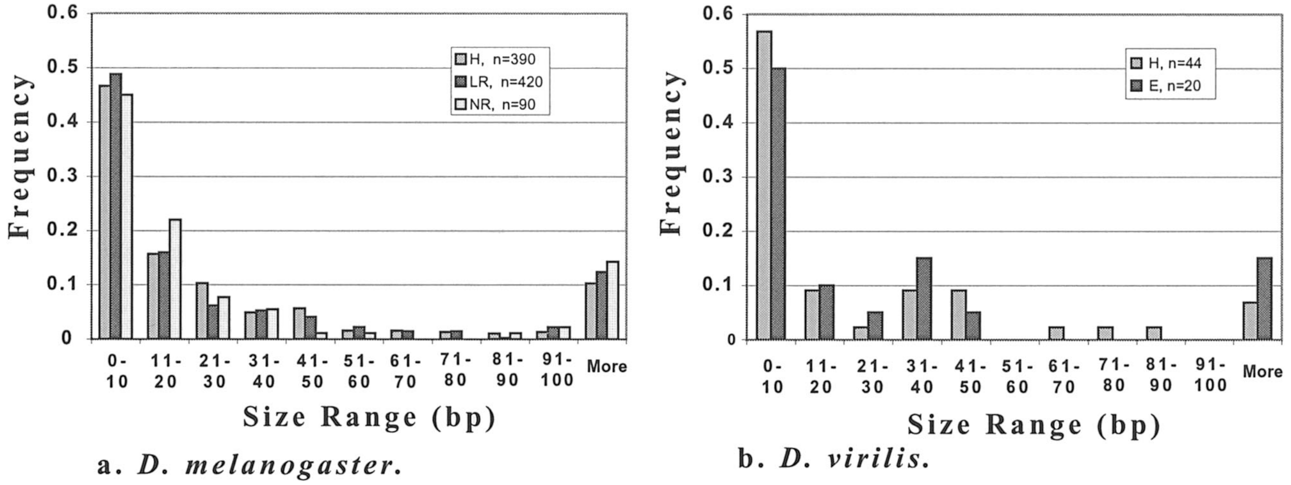
Table 4 shows that mean deletion size is similar in all three regions of the genome. The 95% confidence intervals of the estimates of the mean are substantially overlapping, and means are not significantly different (heterochromatin vs. euchromatin, low recombination: $P = 0.215$; euchromatin, low recombination vs. euchromatin, normal recombination: $P = 0.349$; heterochromatin vs. euchromatin, normal recombination: $P = 0.233$). Because mean deletion size is highly influenced by large outliers, we also determined mean deletion size excluding deletions larger than 400 bp. When these large deletions are excluded, the mean deletion sizes in all regions and for both species are strikingly similar, with overlapping confidence intervals. It is interesting to note that when large deletions are included, the mean deletion sizes, though not significantly different, are larger in regions of euchromatin. This suggests that selection may act to fix very large deletions, but these are much rarer than the smaller common deletions whose fixation does not appear to be influenced strongly by selection. When fragments residing in cytological regions 19 and 20A–B are removed from the analysis of deletion sizes in regions of normal recombination, mean deletion size in regions of normal recombination is 68.7 bp (CI, 15.5 to 166.0 bp). If deletions larger than 400 bp are excluded, mean deletion size in this region is 27.0 bp (CI, 14.1 to 44.0 bp).

Insertion Rates

Using equation (1), we estimated the insertion rate among TEs and genomic regions (tables 5 and 6). Similar to estimates of deletion rates, there is some variation among different TEs. The insertion rate is greatest for the *You* fragment, and the confidence interval does not overlap with those for *jockey, Waldo-A,* and *Helena*. The *X* fragments also show a higher rate of insertion. Insertion rates in *D. virilis Helena* fragments are also slightly elevated relative to those of *jockey, Waldo-A,* and *He-*

Table 3
Deletions per Terminal Branch Substitution for Different Genomic Regions, and 95% Confidence Intervals

Species	Genomic Region	Deletion Rate
D. melanogaster	Heterochromatin	0.119 (0.106 to 0.140)
	Euchromatin, low recombination	0.111 (0.094 to 0.132)
	Euchromatin, normal recombination	0.116 (0.094 to 0.148)
	All regions	0.114 (0.104 to 0.128)
	Petrov and Hartl (1998)	0.120 (0.090 to 0.160)
D. virilis	Heterochromatin	0.161 (0.117 to 0.234)
	Euchromatin	0.254 (0.173 to 0.390)
	All regions	0.174 (0.133 to 0.244)
	Petrov, Lozovskaya, and Hartl (1996)	0.160 (0.090 to 0.260)



a. *D. melanogaster.*

b. *D. virilis.*

FIG. 3.—Frequency distribution of deletion sizes in *D. melanogaster* and *D. virilis* in different genomic domains.

lena fragments in *D. melanogaster*. However, like deletion rates, insertion rates do not differ among different genomic regions in *D. melanogaster*. Estimated rates of insertion are higher in euchromatin versus heterochromatin in *D. virilis*. However, the confidence intervals overlap, and the sample size is very small, so it is difficult to conclude that this finding implies a greater insertion rate in euchromatin in *D. virilis*.

Distribution of Insertion Sizes

Figure 4 shows the distribution of insertion sizes in *D. melanogaster* and *D. virilis* for different genomic regions. Insertions of 1–5 bp make up the vast majority of all insertions. The distribution of insertion sizes is also quite similar among genomic locations. Table 7 shows the mean insertion sizes with 95% confidence intervals. If all insertions are considered, the average insertion size in *D. melanogaster* is highest in heterochromatin and in euchromatic regions with low recombination. These estimates are highly influenced, however, by several large insertions corresponding to TEs and other unidentified sequences repeated in the genome, some of which were greater than 1,000 bp. When such sequences are removed from the analysis, the mean insertion size is similar in all three regions with overlapping confidence intervals. No such repeat sequence insertions were found in *D. virilis*, and the mean insertion size is similar in both genomic regions.

Genome-Wide Rates of DNA Loss in *Drosophila*

Given that there are no significant differences in rates of deletion and insertion among different regions of the *D. melanogaster* genome, it is possible to determine the genome-wide relative rate of deletions and insertions. The genome-wide rate of deletion is 0.114 deletions per terminal branch substitution (CI, 0.104 to 0.128). The genome-wide rate of insertion is 0.032 insertions per terminal branch substitution (CI, 0.025 to 0.04). Thus, deletions are 3.6 times as frequent as insertions, with a 95% confidence interval for this factor of 2.8 to 4.6.

With these estimates it is also possible to calculate the genome-wide rate of spontaneous DNA loss. This is given by the following equation: (average insertion size) × (rate of insertion) − (average deletion size) × (rate of deletion). If we include all deletions and insertions in this rate estimate, DNA is lost at a rate of 5.6 bp per nucleotide substitution. If deletions larger than 400 bp and repeat insertions are excluded, DNA is lost at a rate of 3.6 bp per nucleotide substitution. Using pseudogene data from Graur, Shuali, and Li (1989), the rate of DNA loss in mammals has been estimated to be 0.13 bp per substitution (Petrov and Hartl 1998). Thus, the spontaneous rate of DNA loss is substantially larger in *Drosophila* than in mammals.

Table 4
Mean Deletion Sizes with 95% Confidence Intervals

Species	Genomic Region	Mean Deletion Size (bp, all deletions)	Mean Deletion Size (deletions <400 bp)
D. melanogaster. . .	Heterochromatin	54.2 (39.4 to 70.4)	32.0 (26.7 to 37.5)
	Euchromatin, low recombination	63.0 (47.1 to 81.1)	33.7 (28.1 to 40.4)
	Euchromatin, normal recombination	77.8 (34.9 to 137.5)	33.7 (22.9 to 47.0)
	All regions	60.7 (49.1 to 72.4)	33.0 (29.2 to 36.9)
D. virilis	Heterochromatin	31.1 (15.8 to 49.1)	31.1 (15.8 to 49.1)
	Euchromatin	49.0 (14.6 to 98.6)	29.0 (12.1 to 52.7)
	All regions	36.7 (20.6 to 55.5)	30.5 (18.9 to 44.4)

Table 5
Insertions per Terminal Branch Substitution with 95% Confidence Intervals

Species	Element	Insertion Rate
D. melanogaster	*Waldo-A*	0.022 (0.016 to 0.031)
	Jockey	0.024 (0.018 to 0.032)
	X	0.043 (0.024 to 0.070)
	Helena	0.029 (0.007 to 0.043)
	You	0.075 (0.049 to 0.111)
D. virilis.	*Helena*	0.050 (0.023 to 0.086)

Discussion

Distribution of TEs in the Genome of *D. melanogaster*

Consistent with other studies, we have found that TE density is much higher in heterochromatin and euchromatin with restricted recombination than in euchromatin with normal recombination. Transposable element fragments are about 19-fold denser in heterochromatin than in regions of normal recombination, and mature fragments are about 90-fold denser. There are four potential mechanisms that could explain this difference: (1) there is preferential insertion of TEs into the heterochromatin, (2) there is a lower spontaneous rate of deletion in the heterochromatin, (3) there is a fixation bias in the heterochromatin due to higher selective constraint in other regions of the genome, or (4) there is a fixation bias in the heterochromatin due to positive selection acting to fix TE sequences in these regions.

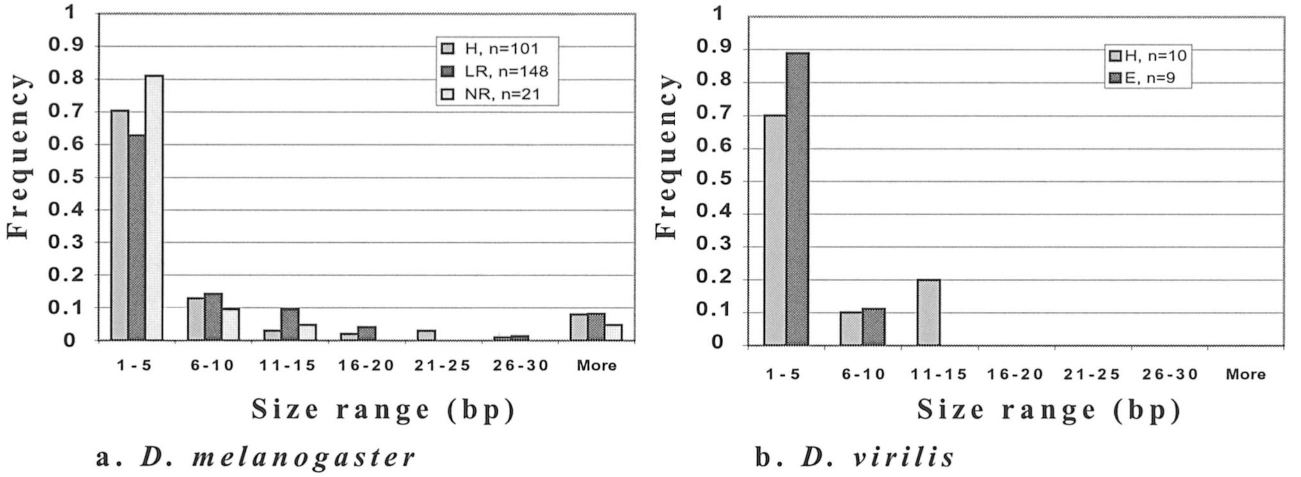
It has been argued that targeted transposition to the heterochromatin can explain the greater non-LTR density in this region (Dimitri and Junakovic 1999). In fact, there is significant evidence that many TEs show preference for certain sites within the genome, such as bent DNA wrapped around nucleosomes (for a review, see Craig 1997). To determine whether transpositional bias could explain the higher density of TEs in heterochromatin, we examined the distribution of fragments ages in different parts of the genome. As can be seen in figure 3, the density of young recently transposed fragments is similar in regions of normal and low recombination, suggesting minimal transpositional bias. Young copies are only about twice as dense in heterochromatin as in regions of normal recombination. Much of this is due to the vastly higher number of young *jockey* fragments that reside in regions of normal recombination (fig. 2*b*). This relationship is less strong for the other elements. However, in the age distribution of all elements (with the exception of *Helena*), the young copies are more proportionally distributed in the three genomic regions than are the older copies. Ignoring the contribution of *jockey* elements, young fragments are about 6.6 times denser in the heterochromatin than in the euchromatic regions of normal recombination. This may reflect a true transpositional bias for these particular elements, although it is possible that selection may have already eliminated many of these sequences from the gene-rich regions of the genome. In any case, it is clear that even a 6.6-fold transpositional bias cannot explain the 90-fold difference in mature TE density between heterochromatin and euchromatin. Although different TEs may have different preferences for regions of the genome, non-LTRs as a group seem to transpose more or less evenly throughout.

For deletions in the size range that we examined, there were no identifiable differences in rates between different regions of the genome. Moreover, the 95% confidence interval for the estimated rate of deletion enables us to reject a rate that is less than 0.106 deletions per substitution in heterochromatin and a rate that is greater than 0.148 deletions per substitution in regions of normal recombination. If we conservatively assume that all TE fragments were inserted at a time corresponding to when the oldest fragment was identified (0.146 terminal branch substitutions/bp) and that a fragment would become unrecognizable in the genome after receiving as many deletions as did the most deleted fragment (0.028 deletions/bp), we can determine the maximum expected difference in density between these two regions that can be explained by differences in the spontaneous deletion rate. Assuming that deletions take place in a Poisson manner, 91% of the fragments in euchromatin would still be detectable if the rate of deletion were 0.148 deletions/bp. By the same token, assuming a deletion rate of 0.106 deletions/bp in heterochromatin, nearly all the fragments (99.9%) would still remain. Under these conservative assumptions, TE fragments would only be expected to be about 9% more frequent in heterochromatin. However, our study has shown that mature TE copies are in fact 90-fold denser in heterochromatin. Thus, differences in spontaneous deletion rate can be eliminated from consideration.

For these reasons we believe that a lower probability of fixation in regions of normal recombination

Table 6
Insertions per Terminal Branch Substitution for Different Genomic Regions, with 95% Confidence Intervals

Species	Genomic Region	Insertion Rate
D. melanogaster	Heterochromatin	0.029 (0.020 to 0.041)
	Euchromatin, low recombination	0.037 (0.026 to 0.050)
	Euchromatin, normal recombination	0.022 (0.015 to 0.038)
	All regions	0.032 (0.025 to 0.040)
D. virilis	Heterochromatin	0.034 (0.013 to 0.056)
	Euchromatin	0.124 (0.050 to 0.276)
	All regions	0.055 (0.027 to 0.094)
Both species.	Petrov and Hartl (1998) estimate	0.015 (0.012 to 0.026)



a. *D. melanogaster*

b. *D. virilis*

FIG. 4.—Frequency distribution of insertion sizes in *D. melanogaster* and *D. virilis* in different genomic domains.

leads to a greater density of older fragments in regions of low recombination. Non-LTR elements land with similar frequency across the genome, as evidenced by similar densities of young elements in the three different regions. However, a greater probability of fixation in regions of low recombination leads to the accumulation of older fragments. This interpretation is supported by our polymorphism study. We showed that five out of five young inserts were polymorphic and five out of five mature inserts were likely fixed in the population. The greater number of older copies in heterochromatin, therefore, seems best explained by the greater likelihood of transition from the polymorphic phase to the fixed phase in this region. One could reasonably expect that the polymorphic young copies, many of which reside in regions of normal recombination, will never become fixed.

What is the nature of selection acting against TEs that can produce such a pattern? It is certainly possible that positive selection may occur in heterochromatin and in regions of low recombination. For example, Dimitri and Junakovic (1999) have suggested that selection for modification of chromatin structure may lead to fixation of repeat sequences in these regions. An important example of the adaptive role that TEs can play is found in *Drosophila* telomeres (Biessmann and Mason 1997). However, because the euchromatic regions of the genome are more gene rich, and also because the deleterious effects of TEs are widely documented, we believe that it is more likely that selection against fixation in euchromatic regions of the genome explains the observed distribution better. This analysis does not distinguish among the proposed models for TE copy number maintenance, such as selection acting against ectopic recombination, as proposed by Montgomery, Charlesworth, and Langley (1987), selection against interference with gene function (Hoogland and Biemont 1996), or a combination of both. Interestingly, this pattern of differential fixation and accumulation has also been noticed in humans. Ovchinnikov, Troxel, and Swergold (2001) have shown that recent LINE-1 insertions are randomly distributed throughout the human genome, whereas older LINE-1 insertions are more common in GC-poor regions. This difference is attributed to different selection pressures for loss or retention in GC-rich and GC-poor regions.

Although many studies have noted an accumulation of TEs in regions of low recombination in *D. melanogaster*, earlier methods of analysis were unable to identify that the discrepancy is primarily due to older sequences, with the exception of those taking a direct cloning approach (Miklos et al. 1988). Earlier methods for determination of chromosomal TE distributions, such as in situ hybridization on polytene chromosomes, were unable to distinguish between young and old copies. Furthermore, PCR-based approaches are unlikely to

Table 7
Mean Insertion Sizes with 95% Confidence Intervals

Species	Genomic Region	Mean Insertion Size (bp)	Mean Insertion Size (repeats removed)
D. melanogaster . . .	Heterochromatin	24.7 (6.6 to 58.9)	5.4 (3.9 to 7.2)
	Euchromatin, low recombination	59.0 (10.4 to 134.6)	5.6 (4.5 to 7.0)
	Euchromatin, normal recombination	6.8 (2.5 to 14.3)	No repeats
	All regions	41.2 (13.0 to 88.7)	5.6 (4.6 to 6.8)
D. virilis	Heterochromatin	3.8 (1.4 to 6.6)	No repeats
	Euchromatin	2.2 (1.0 to 4.2)	No repeats
	All regions	3.2 (1.7 to 4.9)	No repeats

identify old copies in the genome because such sequences will be fairly divergent from the known active TE sequence.

Selection and Deletion in TEs

Analyzing *Helena* DOA copies, Petrov, Lozovskaya, and Hartl (1996) estimated the spontaneous rate of insertion and deletion in *D. virilis* as a function of nucleotide substitution. A very high intrinsic rate of deletion was identified in these fragments, suggesting that *Drosophila* loses DNA at a high rate. However, Charlesworth (1996*a*) argued that selection, either for smaller genome size or reduced ectopic recombination, might cause these estimates to be inflated.

Petrov and Hartl (2000) demonstrated that the selection coefficients expected from selection for small genome size would be too small to cause overestimation of deletion rates. In this article we show that selection for deletions that reduce ectopic recombination also fails to explain the high deletion rate because the estimated deletion rate is indifferent to the local level of recombination. The ectopic model of copy number maintenance provides several predictions. First, if earlier estimates were biased due to selection, it is likely that the fragments previously analyzed were located in regions of effective selection against repeats, namely, in regions of normal recombination; however, if the fragments were located in regions where there is little selection against repeats due to ectopic exchange (i.e., regions of restricted recombination), then the earlier estimates of deletion are unbiased. Second, if selection against ectopic recombination caused rates of deletion to be overestimated, then rates of deletion inferred from regions of normal recombination should be higher than rates inferred from regions of low recombination.

We performed an analysis of rates of deletion in different regions of the genome of *D. melanogaster* to determine whether earlier estimates of rates of deletion were influenced by local effects or inflated due to selection for deletions within TEs. We found that the rate of deletion was very high in heterochromatin and regions of low recombination. We also found no significant difference in the rate of deletion between heterochromatin, regions of euchromatin with low recombination, and regions of euchromatin with normal recombination. Furthermore, we were able to conclude, with 99% confidence, that the rate of deletion in regions of normal recombination could not be more than 32% higher than that in heterochromatin. This provides strong evidence that selection for deletions in TE fragments does not influence our estimates of the rate of spontaneous deletion.

It is widely known that selection is less efficacious in regions of restricted recombination (the Hill-Robertson effect, Hill and Robertson 1966). Thus, if deletions in repeated sequences were positively selected, an increased rate of deletion should be seen within regions of normal recombination. Furthermore, the ectopic recombination model for TE copy number maintenance (Langley et al. 1988) predicts that deletions in repeat sequences would have selection coefficients positively correlated with the rate of recombination because deletions in regions of high recombination would reduce the rate of ectopic exchange more than do deletions in regions of low recombination. Because the effective population size in regions of low recombination is estimated to be 10- to 20-fold smaller than that in regions of higher recombination (Aquadro, Begun, and Kindahl 1994), we can determine an upper bound for the N_es (effective size times selection coefficient) for deletions based on our ability to reject a 32% greater deletion rate in regions of normal recombination relative to heterochromatin. Assuming that the effective population size in heterochromatin is smaller by one order of magnitude, the ratio of fixation probabilities for a new mutation in regions of normal recombination versus one in heterochromatin is given by

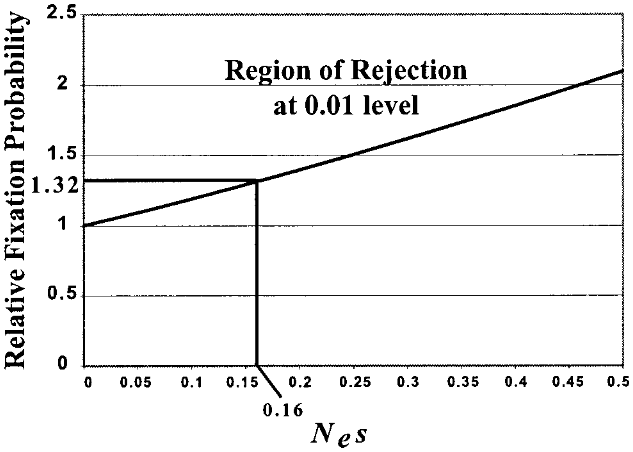


FIG. 5.—The determination of the upper bound of N_es for deletions within non-LTR DOA fragments in *D. melanogaster*. Based on our ability to reject the hypothesis that deletions occur at a 32% greater rate in regions of normal recombination relative to heterochromatin (where rates of recombination are near zero), we are able to reject a value of N_es greater than 0.16. This is based on the assumption that selection is more efficacious in regions of high recombination. The effective population size of *D. melanogaster* has been estimated to be at least 10-fold smaller in regions of low recombination. For a given selection coefficient, equation (3) allows us to determine the relative fixation probabilities for mutations which occur in regions of normal recombination and low recombination.

$$\frac{\dfrac{4N_es}{1 - e^{-4N_es}}}{\dfrac{4\dfrac{N_e}{10}s}{1 - e^{-4(N_e/10)s}}} \tag{3}$$

which is the ratio of fixation probability for a mutation in a population of effective size N_e to that of a mutation in a population with effective size $N_e/10$ (Kimura 1962). Figure 5 shows the ratio of fixation probabilities in regions of normal recombination versus heterochromatin as a function of N_es. As the selection coefficient of deletions in TE fragments increases, the relative fixation frequency of deletions in regions of higher recombination is expected to increase. Because we can reject a rate of deletion difference that is greater than 32% in regions of normal recombination versus heterochromatin, we are also able to reject a value of $N_es > 0.16$ for deletions. This provides a conservative upper bound of N_es for deletions, because the ectopic model of TE copy maintenance predicts that selection coefficients for deletions would be higher in regions of higher recombination. Under these circumstances, an even higher difference in deletion rate would be expected between the two regions if selection were acting to fix deletions.

Given this upper bound on N_es, what can one say about the possible influence that selection may have on estimated rates of spontaneous deletion? An equation from (Kimura 1962):

$$\frac{4N_es}{(1 - e^{-4N_es})} \tag{4}$$

is the probability of fixation of an advantageous allele relative to that of a neutral allele. Given an upper bound of $N_es = 0.16$ for deletions, the relative probability of fixation of an advantageous mutation relative to a neutral one is at most 1.37. Thus, if we have overestimated the rate of spontaneous deletion because of positive selection, the estimate is not likely to be off by more than about 37%. However, the finding of no significant difference as a function of recombination suggests that selection on deletions in this size range is essentially negligible.

We emphasize that our inferences regarding selection apply only to the relatively small deletions in our analysis (predominantly <400 bp). Any analysis of the rates and distributions of deletions must be influenced not only by deletion sizes but also by the sizes of the TE fragments examined. Because in the present study the average fragment size was about 1,600 bp, our study is not informative about the forces acting on deletions larger than 1,600 bp. In fact, in our case only 4% of all deletions were larger than 400 bp. And although we are able to reject the suggestion that selection is acting strongly on the predominant small deletions, there is some evidence that deletions larger than 400 bp may, in fact, be advantageous. Although the differences were not statistically significant, the average deletion sizes in both *Drosophila* species were smallest in heterochromatin, larger in regions of low recombination, and highest in regions of normal recombination. However, when deletions larger than 400 bp were removed from the analysis, the average deletion size was essentially identical in all parts of the genome in both species. This suggests that, unlike prevalent small deletions, rare deletions larger than 400 bp in TE fragments may be positively selected in regions of normal recombination.

Clearly, 400 bp is an arbitrary cutoff between small and large deletions. But in our sample, many of the deletions greater than 400 bp are substantially larger than 400 bp. In our sample, 25% of deletions of >400 bp are larger than 1,000 bp. The average deletion size of deletions <400 bp is 33 bp, whereas the average deletion size of deletions >400 bp is 814 bp. The ectopic model of TE copy number maintenance predicts that the selection coefficient of a deletion should be positively correlated with its size because large deletions in repeated sequences will reduce the likelihood of ectopic recombination to a greater extent than do small deletions. Assuming a linear relationship between the size of the repeated sequence and the likelihood of its participation in ectopic recombination, a 33-bp deletion within a 1,600-bp fragment will reduce the probability of ectopic recombination by only 2%. Deletions between 1 and 10 bp, which comprise 47% percent of all deletions in our analysis, would on average reduce the probability of ectopic recombination only by about 0.3%. On the other hand, an 814-bp deletion will reduce the probability of ectopic recombination by more than 50%. For this reason we believe that predominant small deletions in TE sequences are essentially neutral, but we would not extrapolate this result to larger deletions.

One potential criticism of our attempt to estimate the upper bound of selection coefficients for common deletions is that it relies on estimates of recombination rate. We used frequently reported estimates of the recombination rate in laboratory strains for this analysis. However, there may be a discrepancy between these estimates and those in natural populations. Methods for estimating local recombination rate rely on curve fitting of genetic data from a limited number of crosses. Although these measurements likely represent a broad local average, recombinational hot spots could produce a substantial amount of local variation in recombination rate. Thus, although we estimate an average upper bound for the selection coefficient for small deletions, it is quite likely that this selection coefficient is variable and dependent on many factors. For example, there can be significant variation in the likelihood of participation in ectopic recombination among TEs in similar locations (Montgomery et al. 1991). Because our analysis provides a single estimate for the upper bound of selection coefficients for deletions in repeated sequences, it could be an underestimate if the TE fragments used for this analysis were very different from the average TE. Of course, significant positive selection coefficients (e.g., $N_es > 1$) for deletions will exist only in TE fragments whose initial insertion would be highly deleterious. The fixation of such fragments is unlikely in natural populations.

Future analyses will be needed to determine the variance underlying the estimate of the upper bound for these selection coefficients. This notion does not challenge the major results of this paper: that rates of deletion are very high in regions of restricted recombination and in heterochromatin and that these high rates are probably representative of genome-wide rates of spontaneous deletion. The latter conclusion has important implications for the persistence of gene duplications and the likelihood of the evolution of new gene functions through divergence or subfunctionalization (Lynch and Force 1999; Lynch and Conery 2000).

Deletion Bias in *Drosophila*

We estimate that the ratio of deletions to insertions in the *D. melanogaster* genome is 3.6. This deletion bias is less than that of an earlier estimate of 7.1 in the *D. melanogaster* subgroup based on the *Helena* element alone (Petrov and Hartl 1998). In the present study we found that the ratio of deletions to insertions in *Helena* fragments from the *D. melanogaster* genome is 9.1. The difference between the genome-wide deletion bias estimated using five TE families and that estimated from *Helena* alone suggests that there may be something particular about the *Helena* element. Two possible scenarios could explain this difference. First, selection for elimination could be particularly strong in the *Helena* element, and estimates of the spontaneous rate of deletion based on *Helena* alone could be overestimated. Alternately, there may be something particular about the sequence in the *Helena* element that makes it more prone to deletion. Selection for *Helena* elimination is unlikely as we found no evidence of elevated rates of *Helena* deletion in regions of normal recombination. Therefore, it seems more likely that *Helena* may contain regions that are more deletion-prone.

Although the estimate of the deletion bias is less than that from earlier studies, the estimated overall rates of deletion are very similar. We estimated a genome-wide rate of deletion to be 0.114 deletions per substitution, whereas an earlier estimate of the deletion rate in the *D. melanogaster* subgroup was 0.12 deletions per substitution (Petrov and Hartl 1998). The difference in the estimate of deletion bias can be explained entirely by the somewhat larger insertion rate in the present study. We estimate that the genome-wide rate of insertion is 0.032 insertions per substitution compared with the earlier estimate of 0.017 (Petrov and Hartl 1998). The nearly twofold greater estimate of the insertion rate explains the nearly twofold smaller estimate of the deletion bias. Nevertheless, because our estimate of the average deletion size (excluding deletions greater than 400 bp) is similar to the earlier estimate (33 bp compared with 25 bp), our present data support the contention that *D. melanogaster* has a very high rate of DNA loss due to spontaneous mutation.

Although we have observed a smaller deletion bias than an earlier estimate determined from the *Helena* element alone, our estimate is still substantially larger than that of Comeron and Kreitman (2000). Using polymorphic indel data from introns and flanking regions of genes, they estimated a deletion bias of only 1.35. Furthermore, they estimated a smaller average deletion size (excluding deletions larger than 1,000 bp, the average deletion was 16.3 bp). We believe that selective constraints in introns and flanking regions make these regions problematic for determining the correct distribution and frequency of spontaneous deletions. Many important regulatory motifs reside in these regions, and larger deletions would be expected to have greater deleterious effects than do smaller insertions and deletions. Selection against mutations in these regions is shown by the fact that standing polymorphisms are reduced in these regions relative to synonymous sites (Moriyama and Powell 1996), which themselves experience selection for codon usage bias. For this reason we believe that introns and the flanking regions of genes are poor choices for estimating the rate of spontaneous indels.

The Life History of Different TE Families

By performing an age-distribution analysis on TE fragments, we can learn about their history in their hosts. More than 80% of *jockey* copies in the genome have less than 0.02 terminal branch substitutions/bp. The rest of the *jockey* fragments are much older (0.08–0.16 terminal branch substitutions/bp). This bimodal age distribution of *jockey* elements indicates an ancient mobilization followed by a period of quiescence. Assuming that the rate of nucleotide substitution is 1.5×10^{-8} per year (Rowan and Hunt 1991), the ancient mobilization would have begun approximately 10 MYA—long before the 2- to 3-Myr divergence of *D. melanogaster* and *D. simulans* (Lachaise et al. 1988). The high prevalence of young *jockey* elements indicates that *jockey* TEs are currently active. If these currently active copies fail to fix in the genome due to selection against them, the period of quiescence may be maintained.

The distribution of *You* copies tells a different story of mobilization. There are fewer younger copies of *You* than there are of *jockey,* indicating that it is less currently active. The number of terminal branch substitutions per base pair of all *You* fragments is within the 0.00–0.06 range. Thus, it seems that the wave of mobilization started more recently, approximately 4 MYA, and is still under way.

In sequencing the human genome, Lander et al. (2001) identified different patterns of genome-wide TE mobilization between the ancestors of humans and mice. In the human lineage there seems to have been a peak of mobilization in the past, followed by substantially reduced activity in the last 50 Myr. Transposable element mobilization in the lineage leading to mice, on the other hand, appears to have been constant through time without a recent period of calm. Using the data from whole-genome sequencing projects, one is able to characterize species-specific fluctuations in TE activity. In the future it may be possible to identify the correlation between TE activity and population genetic data. Both sets of data could be used to confirm whether there is increased TE activity during speciation events or population bottlenecks. Furthermore, age analysis of TE fragments should also prove useful in dating events that have been important in shaping genomes.

Acknowledgments

We would like to thank Joshua Cherry for insightful suggestions on the determination of an upper bound for selection coefficients. Lena Hileman provided important suggestions for phylogenetic analysis. We would also like to thank Dan Neafsey, Alex Platt, members of the Hartl and Wakeley laboratories, and our reviewers for helpful comments and discussion. This work was funded by an NIH Genetics Training Grant for J.P.B. and NIH grant GM58423.

LITERATURE CITED

Ananiev, E. V., V. E. Barsky, Y. V. Ilyin, and M. V. Ryzic. 1984. The arrangement of transposable elements in the polytene chromosomes of *Drosophila melanogaster*. Chromosoma **90**:366–377.

Aquadro, C. F., D. J. Begun, and E. C. Kindahl. 1994. Selection, recombination, and DNA polymorphism in *Drosophila*. Pp. 46–56 *in* B. Golding, ed. Non-neutral evolution: theories and molecular data. Chapman & Hall, New York.

Bartolomé, 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.

Berezikov, E. B. A., and I. Busseau. 2000. A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster*. Genome Biol. **1**:1–15.

Biemont, C., A. Tsitrone, C. Vieira, and C. Hoogland. 1997. Transposable element distribution in *Drosophila*. Genetics **147**:1997–1999.

Biessmann, H., and J. M. Mason. 1997. Telomere maintenance without telomerase. Chromosoma **106**:63–69.

Busseau, I., E. Berezikov, and A. Bucheton. 2001. Identification of *Waldo-A* and *Waldo-B*, two closely related non-LTR retrotransposons in *Drosophila*. Mol. Biol. Evol. **18**: 196–205.

Caceres, M., J. M. Ranz, A. Barbadilla, M. Long, and A. Ruiz. 1999. Generation of a widespread *Drosophila* inversion by a transposable element. Science **285**:415–418.

Charlesworth, B. 1996*a*. Genome evolution: the changing sizes of genes. Nature **384**:315–316.

———. 1996*b*. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. **68**:131–149.

Charlesworth, B., P. Jarne, and S. Assimacopoulos. 1994. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. 3. Element abundances in heterochromatin. Genet. Res. **64**:183–197.

Charlesworth, B., C. H. Langley, and P. D. Sniegowski. 1997. Transposable element distributions in *Drosophila*. Genetics **147**:1993–1995.

Charlesworth, B., A. Lapid, and D. Canada. 1992*a*. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. 1. Element frequencies and distribution. Genet. Res. **60**: 103–114.

———. 1992*b*. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. 2. Inferences on the nature of selection against elements. Genet. Res. **60**:115–130.

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. Aguadé. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151**:239–249.

Craig, N. L. 1997. Target site selection in transposition. Annu. Rev. Biochem. **66**:437–474.

Dimitri, P., and N. Junakovic. 1999. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. Trends Genet. **15**: 123–124.

Engels, W. R., and C. R. Preston. 1984. Formation of chromosome rearrangements by *P*-factors in *Drosophila*. Genetics **107**:657–678.

Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics **155**:1405–1413.

Graur, D., Y. Shuali, and W. H. Li. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28**:279–285.

Hartl, D. L. 2000. Molecular melodies in high and low C. Nat. Rev. Genet. **1**:145–149.

Hasegawa, M., H. Kishino, and T. A. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

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. Effect of linkage on limits to artificial selection. Genet. Res. **8**:269–294.

Hoogland, C., and C. Biemont. 1996. Chromosomal distribution of transposable elements in *Drosophila melanogaster*: test of the ectopic recombination model for maintenance of insertion site number. Genetics **144**:197–204.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo, and A. H. Schulman. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA **97**:6603–6607.

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

———. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kindahl, E. C. 1994. Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*. Ph.D. thesis, Cornell University, Ithaca, N.Y.

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

Lachaise, D., M. L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. Historical biogeography of the Drosophila-Melanogaster species subgroup. Evol. Biol. **22**:159–225.

Lander, E. S., L. M. Linton, B. Birren et al. (239 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52**:223–235.

Lim, J. K., and M. J. Simmons. 1994. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays **16**:269–275.

Lozovskaya, E. R., D. Nurminsky, D. A. Petrov, and D. L. Hartl. 1999. Genome size as a mutation-selection-drift process. Genes & Genet. Syst. **74**:201–207.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

Lynch, M., and A. G. Force. 1999. The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. **156**:598–605.

Miklos, G. L. G., M. T. Yamamoto, J. Davies, and V. Pirrotta. 1988. Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the beta heterochromatin of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **85**:2051–2055.

Montgomery, E., B. Charlesworth, and C. H. Langley. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet. Res. **49**:31–41.

Montgomery, E. A., S. M. Huang, C. H. Langley, and B. H. Judd. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. Genetics **129**:1085–1098.

Moriyama, E. N., and J. R. Powell. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**: 261–277.

Myers, E. W., G. G. Sutton, A. L. Delcher et al. (26 coauthors). 2000. A whole-genome assembly of Drosophila. Science **287**:2196–2204.

Nekrutenko, A., and W. H. S. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. **17**:619–621.

Ovchinnikov, I., A. B. Troxel, and G. D. Swergold. 2001. Genomic characterization of recent human *LINE-1* insertions: evidence supporting random insertion. Genome Res. **11**:2050–2058.

Petrov, D. A., and D. L. Hartl. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15**:293–302.

———. 2000. Pseudogene evolution and natural selection for a compact genome. J. Hered. **91**:221–227.

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., J. L. Schutzman, D. L. Hartl, and E. R. Lozovskaya. 1995. Diverse transposable elements are mobilized in hybrid dysgenesis in *Drosophila virilis*. Proc. Natl. Acad. Sci. USA **92**:8050–8054.

Pimpinelli, S., M. Berloco, L. Fanti, P. Dimitri, S. Bonaccorsi, E. Marchetti, R. Caizzi, C. Caggese, and M. Gatti. 1995. Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. Proc. Natl. Acad. Sci. USA **92**:3804–3808.

Priimagi, A. F., L. J. Mizrokhi, and Y. V. Ilyin. 1988. The *Drosophila* mobile element *jockey* belongs to *LINES* and contains coding sequences homologous to some retroviral proteins. Gene **70**:253–262.

Rizzon, C., G. Marais, M. Gouy, and C. Biemont. 2002. Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. Genome Res. **12**:400–407.

Rowan, R. G., and J. A. Hunt. 1991. Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for 5 closely related species of Hawaiian *Drosophila*. Mol. Biol. Evol. **8**:49–70.

SanMiguel, P., A. Tikhonov, Y. K. Jin et al. (8 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science **274**:765–768.

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

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Tatusova, T. A., and T. L. Madden. 1999. BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174**:247–250.

Tudor, T., A. J. Davis, M. Feldman, M. Grammatikaki, and K. O'Hare. 2001. The *X* element, a novel *LINE* transposable element from *Drosophila melanogaster*. Mol. Genet. Genomics **265**:489–496.

Yang, Z. H. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

Antony Dean, reviewing editor

Accepted August 11, 2002