High Rate of DNA Loss in the *Drosophila melanogaster* and *Drosophila virilis* Species Groups

Dmitri A. Petrov[1] *and Daniel L. Hartl*

Department of Organismic and Evolutionary Biology, Harvard University

We recently proposed that patterns of evolution of non-LTR retrotransposable elements can be used to study patterns of spontaneous mutation. Transposition of non-LTR retrotransposable elements commonly results in creation of 5′ truncated, ''dead-on-arrival'' copies. These inactive copies are effectively pseudogenes and, according to the neutral theory, their molecular evolution ought to reflect rates and patterns of spontaneous mutation. Maximum parsimony can be used to separate the evolution of active lineages of a non-LTR element from the fate of the ''dead-on-arrival'' insertions and to directly assess the relative frequencies of different types of spontaneous mutations. We applied this approach using a non-LTR element, *Helena,* in the *Drosophila virilis* group and have demonstrated a surprisingly high incidence of large deletions and the virtual absence of insertions. Based on these results, we suggested that *Drosophila* in general may exhibit a high rate of spontaneous large deletions and have hypothesized that such a high rate of DNA loss may help to explain the puzzling dearth of *bona fide* pseudogenes in *Drosophila.* We also speculated that variation in the rate of spontaneous deletion may contribute to the divergence of genome size in different taxa by affecting the amount of superfluous ''junk'' DNA such as, for example, pseudogenes or long introns. In this paper, we extend our analysis to the *D. melanogaster* subgroup, which last shared a common ancestor with the *D. virilis* group approximately 40 MYA. In a different region of the same transposable element, *Helena,* we demonstrate that inactive copies accumulate deletions in species of the *D. melanogaster* subgroup at a rate very similar to that of the *D. virilis* group. These results strongly suggest that the high rate of DNA loss is a general feature of *Drosophila* and not a peculiar property of a particular stretch of DNA in a particular species group.

Introduction

One of the most striking differences in the genomic organization of *Drosophila* and that of mammals is in the relative abundance of pseudogenes. Pseudogenes are very common in mammals, with many functional genes having more than one pseudogene counterpart and some genes having as many as 200 pseudogene copies (Weiner, Deininger, and Efstratiadis 1986). In contrast, in *Drosophila,* very few pseudogenes have been identified. Two *Drosophila* genes originally identified as pseudogenes (Jeffs and Ashburner 1991; Sullivan et al. 1994) were later shown to be novel functional genes (Long and Langley 1993; Begun 1997).

Beyond the challenge of trying to account for such a striking asymmetry between *Drosophila* and mammals, the dearth of bona fide pseudogenes in *Drosophila* also impedes the study of molecular evolution in this organism. The problem is that in the absence of pseudogenes, it is very hard, or even impossible, to reliably estimate the rates and patterns of spontaneous mutation. This is because mutation is too infrequent to observe directly and, unless one studies neutrally evolving sequences such as pseudogenes, the observed pattern of DNA variation within or between species reflects not only the intrinsic pattern and rate of mutation but also the selective differences between different mutational classes. The problem is particularly acute for the study of indels (insertions and deletions), which are often highly deleterious, and therefore the observed frequencies of indels are undoubtedly profoundly affected by selection.

Although *Drosophila* lacks bona fide pseudogenes, it harbors a class of transposable elements, non-LTR retrotransposons, that commonly generate nonfunctional, pseudogenelike copies as a by-product of the transpositional cycle. We have recently proposed that it is possible to distinguish the evolution of the pseudogenelike copies from the evolution of the transpositionally active lineages, thereby allowing the analysis of the neutral mutational processes without the need for a large number of bona fide pseudogenes.

When an active lineage of a non-LTR retrotransposable element undergoes several independent transpositions, it is expected to produce pseudogenelike copies with essentially identical sequences. After transposition, each one of these copies undergoes independent neutral evolution, which should result in the accumulation of unique point substitutions, deletions, and insertions. If we sample enough independently transposed elements to represent all of the active lineages, we can expect that substitutions that are shared among two or more elements will be those that have occurred in the active lineages themselves. The element-specific substitutions, on the other hand, are those that have accumulated in neutral fashion since the time of transposition, and these substitutions can be used to directly assess the relative frequencies of different types of spontaneous mutations. (For a more detailed account, see Petrov and Hartl [1998].)

We have successfully applied this approach to a non-LTR element, *Helena,* in the *D. virilis* group and, in particular, have demonstrated a surprisingly high rate

[1] Present address: Harvard University Society of Fellows.

Key words: *Helena,* non-LTR retrotransposable elements, biased spontaneous mutation, deletions and insertions, pseudogenes, C-value paradox.

Address for correspondence and reprints: Dmitri Petrov, Harvard University Society of Fellows, 78 Mt. Auburn Street, Cambridge, Massachusetts 02138. E-mail: dpetrov@oeb.harvard.edu.

Mol. Biol. Evol. 15(3):293–302. 1998

of accumulation of relatively large deletions and the virtual absence of insertions (Petrov, Lozovskaya, and Hartl 1996). If these results are general for most of the DNA sequences in *Drosophila,* the fixation of deletions would result in the accelerated loss of superfluous DNA and would also help to explain the observed lack of bona fide pseudogenes in *Drosophila.* We also hypothesized that variations in the rate and the size of indels might contribute to the divergence of genome size in different lineages (the so-called ''C-value'' paradox defined by Thomas 1971) by affecting the amount of ''junk'' DNA in the form of pseudogenes, long introns, intergenic sequences, and so forth.

The validity of these predictions hinges in part on whether the observed high rate of DNA loss in nonfunctional copies of *Helena* is specific to a particular 363 bp of sequence in a particular species group or whether the high rate of deletion is a more general phenomenon. To address these concerns, we have studied the pattern of indel formation and accumulation using a different part of *Helena* in a phylogenetically distant group of *Drosophila,* the *D. melanogaster* species subgroup, which last shared a common ancestor with *D. virilis* 40 MYA (Russo, Takezaki, and Nei 1995). The results presented in this paper demonstrate that the pattern of indel formation is very similar in the *D. melanogaster* and *D. virilis* groups, suggesting that a high rate of DNA loss is common in *Drosophila.*

Materials and Methods

DNA Source

Wild-type stocks of eight species in the *D. melanogaster* species subgroup and a strain of *D. pseudoobscura* were obtained from the *Drosophila* species stock center at Bowling Green, Ohio, or from the species collection in our laboratory. We have utilized the following stocks: *D. erecta* (14021-0224.0), *D. sechellia* (14021-0248.1), *D. orena* (14021-0245.0), *D. pseudoobscura* (14011-0121.0), *D. mauritiana* (w^{pch} strain), *D. simulans* (Congo14), *D. yakuba* (Tai26), *D. teissieri* (Brazzaville8), and *D. melanogaster* (Brazzaville).

Primers, PCR, and DNA Sequencing

Primer sequences were designed using Oligo™ 4.0 (National Biosciences) and supplied by Gibco. PCR reactions were carried out in PTC-100 thermal cyclers (MJ Research) or on a Perkin-Elmer DNA thermal cycler. DNA sequencing was carried out on an ABI373A automated DNA sequencer (Perkin Elmer) with the Taq Cycle Sequencing Dye-Primer or DyeDeoxy Terminator kits from Perkin Elmer. Every nucleotide position was sequenced at least once in both direction.

All clones were obtained by cloning products of PCR reactions carried out with the primers *Helena*96 (5′-CTAAATAACTGCCGAAACAT-3′) and *Helena*1429 (5′-CCTTGCCGTTTGAGTCGTCT-3′). These primers amplify an internal 1,357-bp region of the putative reverse transcriptase gene in *Helena.* The PCR reactions were carried out under the following conditions: 96°C for 30 s, 42°C or 52°C for 1 min, and 72°C for 4 min. In all cases, total genomic DNA was used as template for PCR. Cloning was carried out using the TA-cloning kit (Invitrogen) without prior size fractionation of the DNA. We picked 48 different clones from each PCR reaction for further analysis.

All clones were tested for the presence and the sizes of inserts using PCR with the M13 Universal and Reverse primers. The majority of inserts were of the predicted size of approximately 1,400 bp, with a few clones being somewhat smaller (1,300–700 bp). After sequencing multiple clones (approximately 12 per species), we eliminated all identical sequences, because those most likely correspond to the same genomic insertion of *Helena.* We also eliminated all sequences that did not align with the *Helena* consensus and are likely a result of spurious PCR.

One important concern with using PCR to collect multiple clones for the analysis of length mutations is that the PCR is more efficient in amplifying markedly smaller templates and could possibly bias our sampling procedure. We have implemented a number of steps to minimize any potential bias. First of all, we used long extension times in the initial PCR reaction, which should reduce any preferrential amplification of shorter templates. Cloning and sequencing a large number of clones and using one representative of each unique sequence for further analysis should also reduce the effect of the bias for shorter sequences, because even if a shorter sequence is overrepresented in the initial pool of clones, as long as the longer sequence is present at least once, it will also appear in the final data set. The size distribution of clones after PCR did not reveal any substantial bias for shorter clones. In the actual experiments, most clones were similar in size, ranging from 1,317 to 1,102 bp, making the presence of significant PCR bias unlikely. In addition, we can directly show that shorter clones have not been vastly overrepresented in our sample. For instance, the clone H-*sechellia*455 (1,254 bp) was present twice in our sample, whereas H-*sechellia*468 (752 bp, the shortest sequence in our data set) was cloned only once. Furthermore, the rate of deletions relative to the point substitutions is identical for the 11 shortest sequences in the data set and for the 11 longest ones.

Sequence Analysis

Alignment of all sequences was done with the aid of the MacVector, Genejockey, and Sequencher 2.0 (GeneCodes) software packages. The sequence portions corresponding to the primers *Helena*96 and *Helena*1496 were removed prior to analysis. The phylogenetic analysis used maximum parsimony carried out with the PAUP software package (Swofford 1991). We used all the characters in the nucleotide alignment at equal weight. Deletions were treated as missing data. We also used the MacClade software package (Maddison and Maddison 1992) to aid in tree manipulations. All sequences were deposited in GenBank under the accession numbers AF012030–AF012052. The alignment used in this study can be downloaded from http://www.oeb.harvard.edu/hartl/lab/dmitri.html.

Statistical Methods

Relative rates of deletions versus nucleotide substitutions, and of deletions versus insertions, were estimated using maximum likelihood under the assumptions that (1) each element has no deletions or unique substitutions at the time of transposition, (2) rates of deletions and substitutions are constant in time, and (3) for any given time, the number of deletions and the number of substitutions follow a Poisson distribution. The confidence limits were found using the χ^2 approximation of the log-likelihood ratio. The positive correlation between the number of terminal branch substitutions and the number of deletions was ascertained using Friedman's test for randomized blocks (Sokal and Rohlf 1995).

When considering the rate of deletions per base pair of DNA, one needs to take into account the order in which the deletions have taken place. For instance, suppose that two deletions of 1 and 500 bp occur in a region originally of 1 kb in length. Depending on the order in which the two deletions take place, the estimate of the number of deletions per kilobase can be either 2 (when the 1-bp deletion happens first) or 3 (when the 500-bp deletion is first). This discrepancy is present, because if the 500-bp deletion happens first, then the 1-bp deletion will have taken place in only 500 bp of the remaining DNA, and thus we should count it as 2 events per kilobase. Because we do not know the order of the deletion events in the *Helena* sequence, we have estimated the relative rate of indels using two procedures, one which should underestimate and one which should overestimate the rate. Together, these procedures give us the minimum and maximum values of the rate of deletion. The minimum estimate of deletion rate assumes that all deletions take place simultaneously in a sequence of the original length. To arrive at the maximum estimate, we assume that all deletions take place in a clone of the final length, that is, after the total length of all deletions has been subtracted from the original length of the sequence. The difference between the minimum and maximum estimates in our sample is around 15% of the mean and is well within the 95% confidence interval of either estimate. To be conservative, and also for the sake of brevity, we report only minimum estimates of the deletion rate.

The calculation of a half-life of a pseudogene was done using a continuous decay formula: $L = L_0\exp(-\mathrm{rt})$, where L is the length of a pseudogene at time t, L_0 is the length at time 0, and r is the deletion rate (product of the average size of a deletion by the rate of deletions per substitution or per year). With the rate of DNA loss given in Myr, we used the following estimates of the neutral substitution rates: 5×10^{-9} substitutions/year for mammals and 15×10^{-9} substitutions/year for *Drosophila* (Sharp and Li 1989).

Results and Discussion

Sampling and Phylogenetic Analysis of *Helena* in the *D. melanogaster* Subgroup

Helena, a non-LTR retrotransposable element that was originally identified in *D. virilis* (Petrov et al. 1995), is widely distributed in the genus *Drosophila* (unpublished data) and, in particular, is present in all species of the *D. melanogaster* subgroup as well as in *D. pseudoobscura.* We have cloned and sequenced multiple insertions of *Helena* in all eight species of the *D. melanogaster* subgroup (*D. orena, D. erecta, D. teissieri, D. yakuba, D. melanogaster, D. sechellia, D. simulans,* and *D. mauritiana*) and have obtained a single clone from *D. pseudoobscura.* The cloning procedure involved carrying out PCR reactions with two primers designed to amplify the internal 1,357-bp region in the putative reverse transcriptase gene of *Helena* and then cloning the products of the reaction and sequencing individual clones. In an effort to sample independently transposed elements, we used a single strain per species as template DNA for PCR, which should reduce the probability of vertical transmission and of resampling of the elements that are present at the same site in the genome (Petrov and Hartl 1998). On the other hand, because each species carries multiple insertions of *Helena,* this procedure is likely to result in sampling of the same insertion in more than one clone. To minimize this problem, we excluded all identical sequences from our analysis. In this way, we obtained 23 different sequences from eight species in the *D. melanogaster* subgroup and 1 sequence from *D. pseudoobscura.*

The alignment of the sampled sequences revealed a large number of indels, including 64 apparent deletions and 8 apparent insertions (fig. 1). However, outside of the indels, the alignment was unambiguous. To further investigate the evolution of *Helena* in the *D. melanogaster* subgroup, we performed phylogenetic analysis with all sampled sequences using maximum parsimony. The resulting tree, in figure 1, is the strict consensus of 12 equally parsimonious trees.

Evolutionary History of *Helena* in the *D. melanogaster* Subgroup

The evolutionary history of each independently transposed non-LTR element can be separated into two distinct phases: (1) the evolution of an active lineage, which is reflected in the sequence of each element at the time of transposition, and (2) the pseudogenelike neutral evolution of each element after transposition. Because multiple independently transposed elements generated from the same active lineage are expected to have the same sequence at the time of transposition, it follows that substitutions that occur in active lineages must be shared among several elements. On the other hand, the pseudogenelike evolution of each element after transposition should result in unique substitutions. Based on this reasoning, we have argued in our previous reports (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998) that maximum parsimony should distinguish between these two phases, such that the evolution of the active lineages will be represented by substitutions that are shared among several elements (these are changes that map to the internal branches), whereas neutral drift will be reflected in the element-specific terminal branches. The basis of this separation is valid as long as all of the active lineages are represented in our sample more

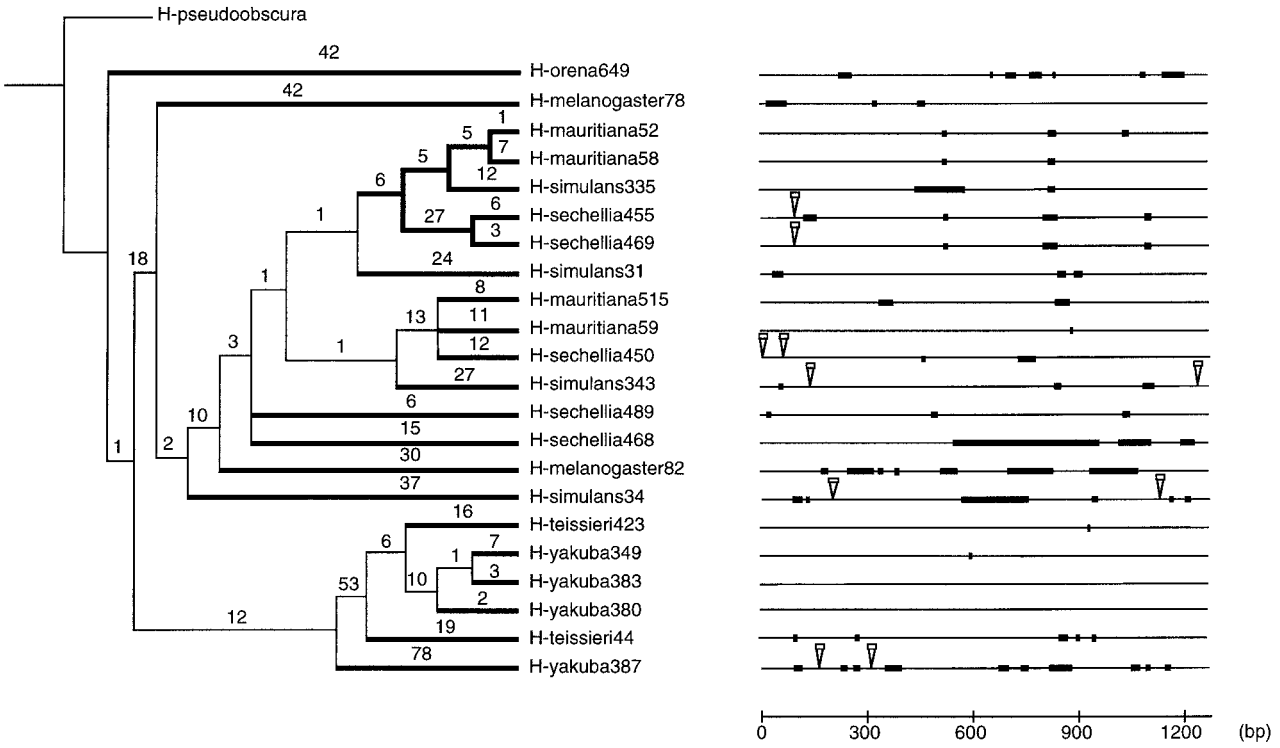

FIG. 1.—Phylogenetic analysis and the locations and sizes of deletions in copies of *Helena* from the *D. melanogaster* species subgroup. Schematic diagram of the location of deletions in the 22 aligned sequences of *Helena* are shown on the right. A sequence of *Helena* from *D. pseudoobscura* was used to root the tree. Maximum-parsimony analysis was carried out using all positions in the nucleotide alignment at equal weight. We have ignored the insertions and have treated deletions as ''missing data.'' The number of unambiguous substitutions is shown above each branch. Deletions are shown as filled-in black bars, with the length of each bar corresponding to the length of each deletion. Insertions are represented by triangles. The bold lines indentify the ''pseudogene'' branches (see text for explanation).

than once, as long as all the terminal sequences correspond to independently transposed elements, and as long as inactive elements cannot transpose in trans. (For a more detailed account, see Petrov and Hartl [1998].)

Thus we expect to find evidence of purifying selection acting along the internal branches of the tree and evidence of a lack of constraint on the terminal branches. The distribution of point substitutions along the branches of the tree fits this expectation (fig. 2*a*): in the internal branches there is a sharp excess of third-position substitutions, indicating the action of purifying selection against amino acid substitutions ($\chi^2 = 66.6$, $P = 3.5 \times 10^{-15}$), whereas in the terminal branches, point substitutions are distributed evenly among all codon positions ($\chi^2 = 0.54$, $P = 0.76$). In one respect, however, the distribution of indels and stop codons is not entirely consistent with our expectations. Because most indels and stop codons in the coding regions are likely to abolish the activity of reverse transcriptase, we expect that indels and stop codons should appear exclusively in the terminal branches. The vast majority of indels and stop codons (60 of 64 deletions, 7 of 8 insertions, and 17 of 18 stop codons) do map to the terminal branches. However, seven deletions, one insertion, and one stop codon are shared among two to five different sequences.

The shared indels and stop codons can be mapped onto the internal branches in a way that is completely consistent with the tree without invoking parallel changes or reversals. A key observation is that all of them map to the branches connecting five elements, namely *mauritiana*52, *mauritiana*58, *simulans*355, *sechellia*455, and *sechellia*469 (fig. 3). The presence of indels and stop codons in the internal branches may imply that these particular changes did not significantly interfere with the transpositional competence of the active lineage. This possibility implies that in the branches shared between the elements *sechellia*455 and *sechellia*469, three deletions (of 1, 4, and 24 bp) and one insertion (of 1 bp), which together remove 28 bp of DNA and result in a frameshift, nevertheless allowed two independent transpositions that produced *sechellia*455 and *sechellia*469. Alternatively, it is possible that five of the inconsistent elements correspond to a single transpositional event, with subsequent vertical transmission and resampling of the same element in five different cases.

There are a number of reasons why we favor the resampling hypothesis. First, all of the indels and stop codons are concentrated in one part of the tree. If some indels and stop codons were consistent with transposition in general, we would expect to see them in other parts of the tree as well. Second, all five of the elements were sampled from a cluster of the very closely related species *D. mauritiana*, *D. simulans*, and *D. sechellia* (Lemeunier and Ashburner 1976), which makes sampling of a vertically transmitted inactive element a more likely scenario. Finally, vertical transmission of a single insertion should lead to a general release from purifying selection along the part of the tree that connects these

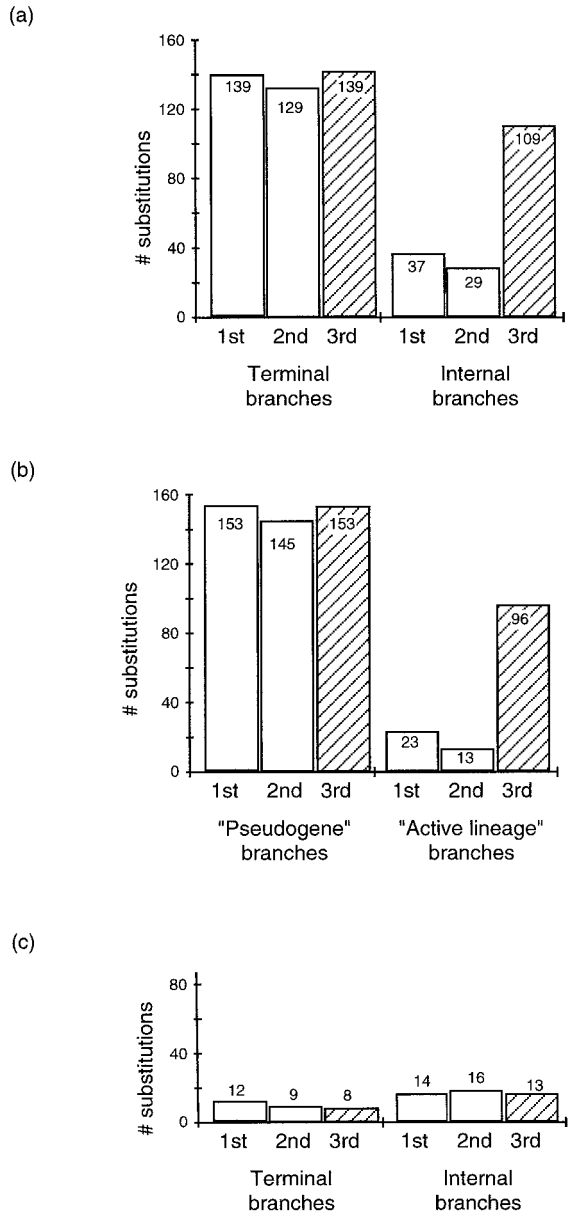

FIG. 2.—Distribution of substitutions by codon position along branches of the *Helena* gene tree. *a,* Distribution of substitutions along the terminal and internal branches of the *Helena* gene tree. The internal branches show clear signs of purifying selection ($P = 3.5 \times 10^{-15}$), whereas the terminal branches do not ($P = 0.76$). *b,* Distribution of substitutions along the ''pseudogene'' and ''active'' branches of the tree. The ''pseudogene'' branches comprise all terminal branches combined with the branches of the *mauritiana*52–*sechellia*469 clade (see text and fig. 3). The ''active'' branches comprise all of the internal branches with the exclusion of the *mauritiana*52–*sechellia*469 clade. Purifying selection is even more pronounced in the ''active'' branches ($P = 5.4 \times 10^{-21}$) than in the internal branches and is absent in the ''pseudogene'' branches ($P = 0.86$). *c,* Lack of purifying selection along the internal branches ($P = 0.85$) and the terminal branches ($P = 0.63$) of the *mauritiana*52–*sechellia*469 clade.

five elements. This prediction is consistent with the presence of deletions and insertions, and it is also supported by the distribution of point substitutions along the internal branches connecting these elements (fig. 2*c*): the point substitutions are distributed evenly among the three codon positions ($\chi^2 = 0.32$, $P = 0.85$).

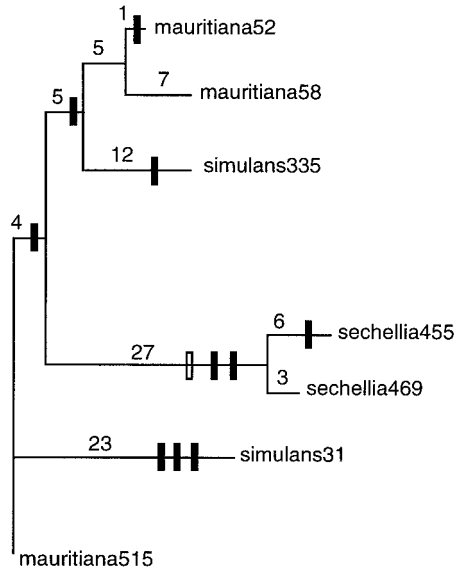

FIG. 3.—The *mauritiana*52–*sechellia*469 clade, which has all of the shared deletions and insertions. Deletions are represented by filled-in bars, insertions by open bars. Numbers of point substitutions mapping to each branch are indicated above each branch.

We therefore conclude that the internal branches leading to the five inconsistent elements appear to correspond to the vertical evolution of a single insertion in the common ancestor of *D. mauritiana, D. sechellia,* and *D. simulans.* These branches are therefore combined with all of the terminal branches to arrive at the set of ''pseudogene'' branches that correspond to the pseudogenelike part of evolution in our sample of *Helena* insertions. (These ''pseudogene'' branches are represented by bold lines in fig. 1.)

The Numbers of Deletions and Point Substitutions Are Positively Correlated Along the ''Pseudogene'' Branches

The number of deletions, insertions, and point substitutions along each ''pseudogene'' branch should be proportional to the amount of time that has elapsed for each element after its transposition. We therefore expect to find a positive correlation between the numbers of any two types of substitutions for any ''pseudogene'' branch. We do find such a correlation for the numbers of deletions and point substitutions (fig. 4; $P = 0.008$, Friedman's method for randomized blocks). We cannot, however, demonstrate a correlation between the number of insertions and the number of point substitutions ($P = 0.38$, Friedman's method for randomized blocks), which is probably due to the small number of insertions (8) in the sample and, consequently, insufficient power to detect a correlation. Visual inspection of figure 1 shows that the distribution of insertions is at least consistent with the model. Most insertions are present on the longest branches that also have a large number of substitutions and deletions (*yakuba*387, *simulans*34); and the shortest branches are free of insertions (*mauritiana*52, *yakuba*383). In fact, the branches with insertions have, on average, 2.8 times more point substitutions and 2.3 times more deletions than branches without insertions.

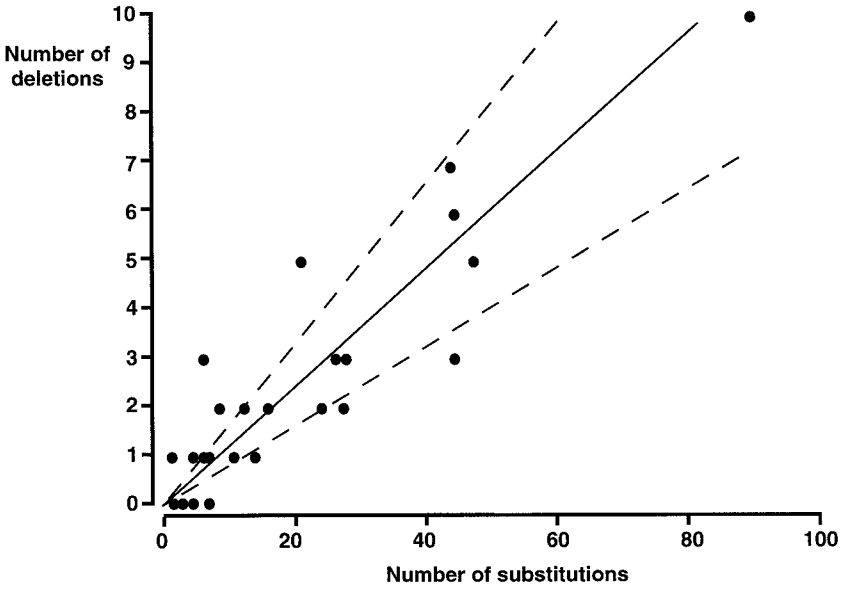

FIG. 4.—The solid line shows the maximum-likelihood regression between the numbers of deletions and point substitutions along the ''pseudogene'' branches. The dashed lines represent the 95% confidence interval of the rate of deletions relative to the number of substitutions.

The presence of the positive correlation also allows us to estimate the relative mutation rates of deletions and point substitutions. After correcting the number of substitutions for the length of each copy of *Helena* sequences and for multiple hits using the one-parameter Jukes-Cantor method, we arrive at the maximum-likelihood estimate of 0.12 deletions per substitution (the 95% confidence interval is 0.09–0.16). Note that this is a conservative minimum estimate of the rate of deletions (see *Materials and Methods*). This estimate of 0.12 deletions per substitution is marginally smaller than the one we reported for the *D. virilis* group (0.16 deletions per substitution), but it is not significantly different (*G*-test, $P = 0.47$). The combined rate of deletion in these two groups is 0.13 deletions per substitution, with a 95% confidence interval of 0.12–0.14 deletions per substitution.

Pattern of Indels in the *D. melanogaster* and *D. virilis* Species Groups

Deletions in the *D. melanogaster* subgroup *Helena* sample range from 1 to 432 bp, with a mean of 34 bp and a standard deviation of 65 bp. The distribution of deletion sizes is highly asymmetrical: 62% of all deletions range from 1 to 20 bp, 19% range from 21 to 50 bp, and 19% range from 51 to 432 bp. Deletions of 1 bp are the most frequent; they account for 26% (17 of 64) of all deletions.

The lengths of the sequenced portions of *Helena* in the *D. melanogaster* and *D. virilis* data sets are different (1,317 bp compared to 363 bp), which prevents us from comparing average lengths of deletions directly. The problem is that we would expect to miss most of the large deletions in the shorter sequences in the *D. virilis* data set, because it is impossible to observe a deletion larger than 363 bp in a sequence only 363 bp in length. In order to make the distribution of deletion sizes in the *D. virilis* and *D. melanogaster* data sets commensurable, we used a sliding window of 363 bp and a step length of 50 bp to extract all the deletions in the *D. melanogaster* data set that have both of their breakpoints inside this window. The resulting simulated distribution of deletions is that expected from a sequenced region of *Helena* of 363 bp instead of one of 1,350 bp in species of the *D. melanogaster* subgroup. The average size of deletions in the simulated *D. melanogaster* data set is 25.0 bp, which is very close to that observed in species of the *D. virilis* group (24.3 bp). We can also compare the shapes of the distribution of deletion size in the *D. virilis* group and in the simulated *D. melanogaster* data set, which appear very similar (fig. 5).

In contrast to the large number and the average size of deletions, we observed only eight small insertions. The insertions range from 1 to 7 bp and average 2.8 ± 2.3 bp. The only insertion in the *D. virilis* data set is a tandem duplication of 4 bp that falls in the middle of the distribution of insertion sizes in the *D. melanogaster* data set. However, given the extremely small number of insertion events, no meaningful comparison of the insertion sizes is possible.

We can also compare the relative frequencies of deletions and insertions in the *D. melanogaster* subgroup (64 deletions vs. 9 insertions) and in the *D. virilis* group (23 deletions vs. 1 insertion). A *G*-test with the Yates correction for the small number of insertions fails to reveal significant heterogeneity ($P = 0.57$).

We have previously reported that about half of the deletions in the *D. virilis* data set can be inferred to have been flanked by short 2–7-bp direct repeats, one of which is deleted along with the intervening sequence (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998); the only insertion in the *D. virilis Helena* data is a tandem duplication of 4 bp. We observe an essentially identical pattern in the *D. melanogaster* data set (data not shown). Approximately half of all deletions are

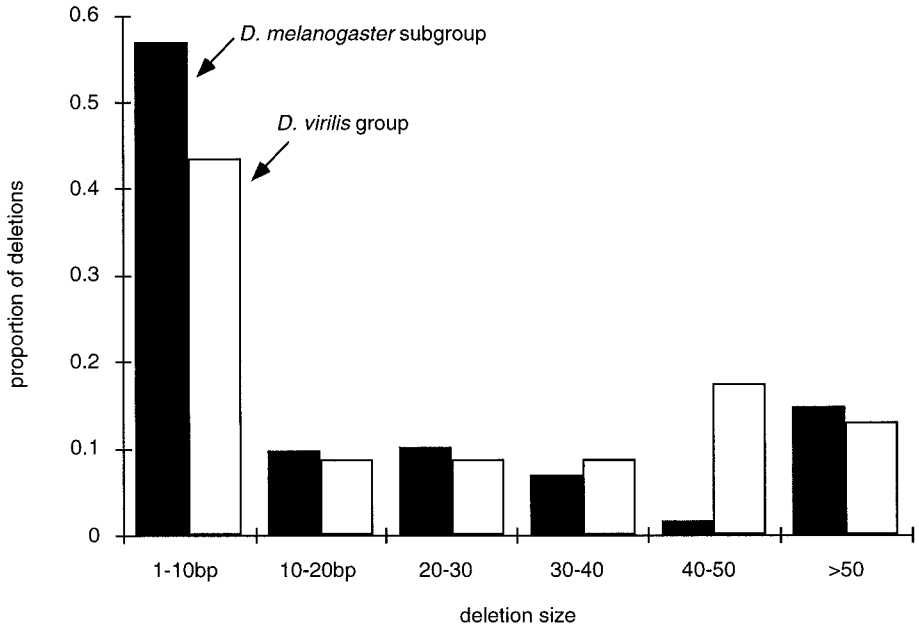

FIG. 5.—Distribution of deletion sizes in the *D. virilis* (Petrov, Lozovskaya, and Hartl 1996) and in the simulated *D. melanogaster* data sets (see text for explanation). Each bar represents the proportion of deletions in the data set that fall within a particular size range. The *D. melanogaster* deletions are represented by black bars, and the *D. virilis* deletions are represented by white bars.

flanked by direct duplications of 1–7 bp, and 6 of 9 insertions are tandem duplications of 1–7 bp. The similarity of these pattern suggests that indels in the *D. melanogaster* and *D. virilis* species groups are generated by similar mechanisms. Combined with similar size distributions and rates of formation of indels in the two groups, it appears that the patterns of indel evolution in the *D. virilis* and *D. melanogaster* groups are indistinguishable.

Is the Apparent Size Distribution of Indels Biased by Selection for Smaller Genome Size?

The central assumption of our study is that individual insertions of non-LTR elements evolve neutrally and accumulate point substitutions, deletions, and insertions in proportion to the likelihood of their spontaneous formation. We have supported this claim by demonstrating that point substitutions along ''pseudogene'' branches map to the first, second, and third positions of codons with equal probability, signifying a lack of purifying selection on this sequence for the ability to produce the functional reverse transcriptase. Because we sequenced a part of the coding region and not a regulatory sequence, such as an enhancer, a silencer, or a binding site for a chromatin-related protein, it is also unlikely that the mere presence of this sequence would have any direct biological activity and thus be affected by selection. These considerations make it seem likely that the pattern of point substitutions along ''pseudogene'' branches in our data set does indeed reflect the neutral pattern and rate of mutation.

The situation is potentially more complicated in the case of indels. Indels do not merely affect the function of the gene in which they occur, they may also have more global effects by, for example, changing the total length of the genome. Differences in the total amount of DNA can lead to variation in time of replication, energy required for proper chromatin packaging, and so forth, all of which can be nonneutral. If selection favors smaller genome size, that in itself might bias our data set toward larger deletions and against insertions of any size. The greater efficacy of selection in *Drosophila* owing to much larger population size might then account for the discrepancy in average deletion size between *Drosophila* and mammals (Charlesworth 1996; Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998).

One prediction of this ''selectionist'' model would be that selection would tend to eliminate insertions of individual elements in proportion to the total amount of DNA that they add to the genome. Elements that accumulate more and, importantly, longer deletions will be more likely to persist in populations for longer periods. We would therefore expect to observe a positive correlation between the ages of individual elements and the total number and the lengths of deletions. The age of a pseudogenelike element is proportional to the number of point substitutions accumulated since transposition. The ''neutralist'' model, on the other hand, predicts a positive correlation only between the number of point substitution and the number of deletions, not between the lengths of the deletions and the number of point substitutions. Both deletions and point substitutions should accumulate with time, but long and short deletions should be observed in young and old elements with equal likelihood.

As predicted by both models, numbers of deletions and point substitutions do correlate in the *Helena* data sets in both the *D. virilis* group (Petrov et al. 1996; Petrov and Hartl 1998) and the *D. melanogaster* sub-

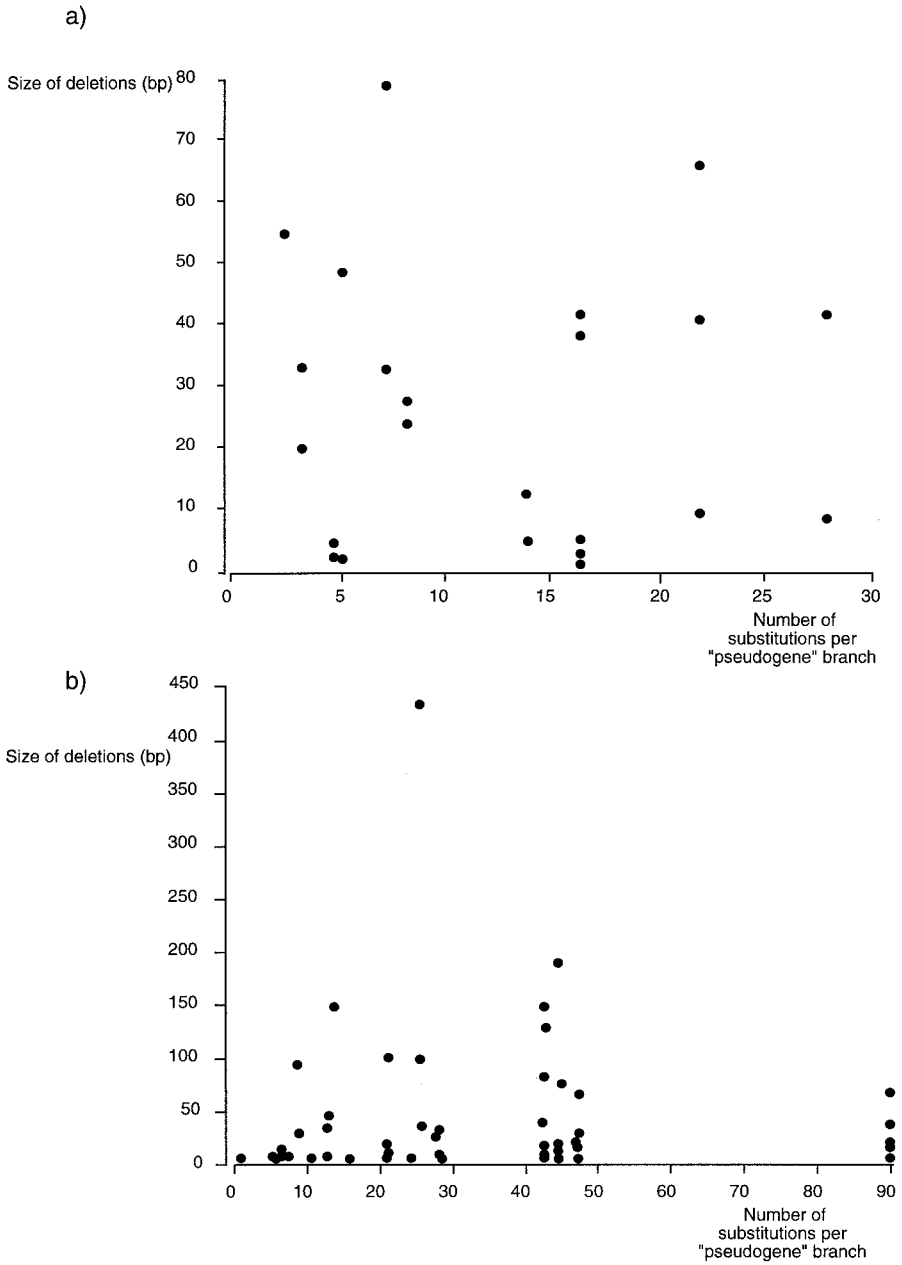

FIG. 6.—*a*, Lack of correlation between the number of point substitutions and the sizes of deletions in individual elements in the *D. virilis* data set (Petrov, Lozovskaya, and Hartl 1996) (Friedman's method for randomized blocks, $P = 0.3$). *b*, Lack of correlation between the number of point substitutions and the sizes of deletions in the *D. melanogaster* data set (Friedman's method for randomized blocks, $P = 0.9$).

group (fig. 4; $P = 0.008$, Friedman's method for randomized blocks). Neither data set, however, shows signs of a positive correlation between the lengths of deletions and the number of point substitutions (fig. 6). The absence of a detectable bias toward longer deletions in older sequences argues that the observed pattern and rate of indels in *Drosophila* is primarily the product of spontaneous mutation biased toward frequent long deletions and rare short insertions. Selection for smaller genome size may indeed be operating in *Drosophila*, but it is apparently not very efficacious when applied to indels of one to a few hundred base pairs.

High Rate of DNA Loss in Both the *D. melanogaster* and *D. virilis* Species Groups

We have previously suggested that *Drosophila* exhibit a high rate of DNA loss through the biased accumulation of large deletions (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998). We have based this suggestion on the demonstration that a 363-bp region of the reverse transcriptase gene in a non-LTR element *Helena* in the *D. virilis* group preferentially accumulates large deletions when the gene is relieved of functional constraints. The validity of this claim depends on the assumptions that (1) the analyzed 363-bp region of *Hel-*

Table 1
Indel Evolution in Mammals and *Drosophila*

	Drosophila	Mammals[a]	Significance of Difference
Ratio of insertions to point substitutions	0.015 (0.012–0.026)[b]	0.010 (0.006–0.013)	NS[c]
Ratio of deletions to insertions	8.7 (6.2–17.8)	4.7 (3.1–7.3)	NS
Ratio of deletions to point substitutions	0.13 (0.12–0.14)	0.049 (0.041–0.058)	$P \ll 0.05$
Mean deletion size (bp)	24.9 ± 37.0	3.2 ± 4.6	$P \ll 0.05$
Mean insertion size (bp)	2.9 ± 2.3	2.4 ± 2.1 (8.5 ± 24.1)[d]	NS
Half-life of a pseudogene (point substitutions per nucleotide)	0.21	4.42	$P \ll 0.05$
Half-life of a pseudogene (Myr)	14.3	884	$P \ll 0.05$

[a] Data are from Graur, Shuali, and Li (1989).
[b] 95% confidence interval of the estimate.
[c] Not statistically significant.
[d] Taking into account a single 125-bp insertion in the rat a-tubulin.

ena is an unbiased representative of most of the DNA sequences in *Drosophila*, and (2) the mutational pattern of indels in the *D. virilis* group is not significantly different from that in *Drosophila* in general.

To test these assumptions, we analyzed the pattern of indels using a different part of *Helena* in the *D. melanogaster* subgroup, which is distantly related to *D. virilis*. The main conclusion is that, in all respects, the patterns of spontaneous formation of indels in these two groups are indistinguishable. We did not detect any significant differences in the relative frequencies of deletions and insertions, the relative rates of deletions and point substitutions, or the size distributions of indels. In addition, indels in both groups are likely to be formed by similar mechanisms, as indicated by the presence of short direct repeats flanking many of the deletions in both data sets. The fact that we have observed such similar patterns of indel formation in two unrelated sequences boosts our confidence that this pattern is general for a large proportion of sequences in the *Drosophila* genome.

Drosophila melanogaster and *D. virilis* belong to different subgenera of *Drosophila*, *Sophophora* and *Drosophila*, respectively. They last shared a common ancestor approximately 40 MYA and represent one of the deepest splits in the drosophilid phylogeny (Russo, Takezaki, and Nei 1995). Hence the similarity of the patterns of indel formation argues strongly that the high rate of DNA loss is prevalent and probably ancestral for all drosophilids.

Because the *D. melanogaster* and *D. virilis* data sets are so similar, we can combine them to more accurately compare indel evolution in *Drosophila* and mammals (Graur, Shuali, and Li 1989). Table 1 summarizes these comparisons. For insertions, there is no profound difference between *Drosophila* and mammals. Both the relative rates of insertions compared to point substitutions, and the average sizes of insertions are very similar. Deletions, on the other hand, are both more prevalent and much larger in *Drosophila* than in mammals. Because insertions are so infrequent and short compared to deletions, we ignored them in our estimation of the rate of DNA loss.

The relative rate of deletions per point substitution is 2.6 times higher in *Drosophila* than in mammals. Even more pronounced is the difference in the average sizes of deletions, which are almost eight times larger in *Drosophila*. The higher rate of formation and the larger average size of deletions combine to eliminate DNA approximately 20-fold faster in *Drosophila* than in mammals. Taking into account that the rate of point substitutions is about threefold higher in *Drosophila* than in mammals (Sharp and Li 1989), we estimate that *Drosophila* loses nonessential DNA at a rate that is approximately 60 times higher than that in mammals. Thus, a pseudogene fixed in a mammalian lineage is expected on average to lose half of its DNA in 884 Myr—an extremely long period, even on an evolutionary timescale—and it will become unrecognizable owing to point substitutions long before then. In contrast, a *Drosophila* pseudogene is expected to lose half of its DNA in only 14.3 Myr. To put this in the context of *Drosophila* evolution, the evolutionary distance between *D. melanogaster* and *D. yakuba* is approximately 12 Myr, so homologous pseudogenes in *D. melanogaster* and *D. yakuba* will share only 56% of their DNA and would be unlikely to either cross hybridize or even be alignable should they be sequenced. If the rates of pseudogene formation in mammals and *Drosophila* are similar, the higher rate of DNA elimination will significantly reduce the probability of observing a pseudogene in *Drosophila* at any given time.

Variation in the rate of DNA loss among different lineages may also contribute to the differences in genome size by affecting the amount of superfluous DNA in the form of pseudogenes, long introns, intergenic regions, and so forth. If this is true, then lineages with high rates of DNA loss should have small, ''tidy'' genomes with few pseudogenes and short introns, whereas lineages with low rates of DNA loss should have large, ''messy'' genomes with large proportions of ''junk'' DNA of all kinds. The absence of research that combines measurements of the amount of ''junk'' DNA, genome size, and estimates of the rate of spontaneous DNA loss due to biased mutation in different lineages precludes immediate evaluation of this hypothesis. The simplicity of estimating relative rates and size distributions of deletions and insertions using non-LTR retrotransposable elements (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998), combined with their ex-

tremely wide phylogenetic distribution (Kimmel, Ole-Moiyoi, and Young 1987; Schwarz-Sommer et al. 1987; Finnegan 1989*a*, 1989*b*; Hutchison et al. 1989; Cambareri, Helber, and Kinsey 1994), should prove useful in resolving these kinds of issues.

Acknowledgments

We thank M. Siegal, D. Weinrech, R. Lewontin, P. Goss, and members of our laboratory for helpful discussions. Comments by C. Aquadro and two anonymous reviewers substantially improved the manuscript. This work was supported by NIH grants GM33741 and HG01250.

LITERATURE CITED

BEGUN, D. 1997. Origin and evolution of a new gene descended from *alcohol dehydrogenase* in Drosophila. Genetics. **145**:375–382.

CAMBARERI, E. B., J. HELBER, and J. A. KINSEY. 1994. Tad1-1, an active LINE-like element of *Neurospora crassa.* Mol. Gen. Genet. **242**:658–665.

CHARLESWORTH, B. 1996. The changing sizes of genes. Nature. **384**:315–316.

FINNEGAN, D. J. 1989*a*. *F* and related elements in *Drosophila melanogaster*. Pp. 519–522 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology, Washington, D.C.

———. 1989*b*. The *I* factor and I-R hybrid dysgenesis in *Drosophila melanogaster*. Pp. 503–518 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology, Washington, D.C.

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.

HUTCHISON, C. A. III, S. C. HARDIES, D. D. LOEB, W. R. SHEHEE, and M. H. EDGELL. 1989. LINEs and related retroposons: long interspersed repeated sequences in the eukaryotic genome. Pp. 593–618 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology, Washington, D.C.

JEFFS, P., and M. ASHBURNER. 1991. Processed pseudogenes in *Drosophila.* Proc. R. Soc. Lond. B. **244**:151–159.

KIMMEL, B. E., O. K. OLE-MOIYOI, and J. R. YOUNG. 1987. Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol. Cell. Biol. **7**:1465–1475.

LEMEUNIER, F., and M. ASHBURNER. 1976. Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*Sophophora*). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond. B **193**:275–294.

LONG, M. Y., and C. H. LANGLEY. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila.* Science **260**:91–95.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade. Version 3. Sinauer, Sunderland, Mass.

PETROV, D. A., and D. L. HARTL. 1998. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila.* Gene (in press).

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.

RUSSO, C. A. M., N. TAKEZAKI, and M. NEI. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. **12**:391–404.

SCHWARZ-SOMMER, Z., L. LECLERCQ, E. GOBEL, and H. SAEDLER. 1987. Cin4, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J. **6**:3873–3880.

SHARP, P. M., and W.-H. LI. 1989. On the rate of DNA sequence evolution in *Drosophila.* J. Mol. Evol. **28**:398–402.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry: the principles and practice of statistics in biological research. W. H. Freeman, New York.

SULLIVAN, D. T., W. T. STARMER, S. W. CURTISS, M. MENOTTI-RAYMOND, and J. YUM. 1994. Unusual molecular evolution of an *Adh* pseudogene in *Drosophila.* Mol. Biol. Evol. **11**: 443–458.

SWOFFORD, D. L. 1991. PAUP: phylogenetic analysis using parsimony. Version 3.0s. Illinois Natural History Survey, Champaign.

THOMAS, C. A. 1971. The genetic organization of chromosomes. Annu. Rev. Genet. **5**:237–256.

WEINER, A. M., P. L. DEININGER, and A. EFSTRATIADIS. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.

CHARLES F. AQUADRO, reviewing editor

Accepted November 10, 1997