Rapid Sequence Turnover at an Intergenic Locus in Drosophila

Nadia D. Singh and Dmitri A. Petrov

Department of Biological Sciences, Stanford University, Stanford, California

Closely related species of Drosophila tend to have similar genome sizes. The strong imbalance in favor of small deletions relative to insertions implies that the unconstrained DNA in Drosophila is unlikely to be passively inherited from even closely related ancestors, and yet most DNA in Drosophila genomes is intergenic and potentially unconstrained. In an attempt to investigate the maintenance of this intergenic DNA, we studied the evolution of an intergenic locus on the fourth chromosome of the *Drosophila melanogaster* genome. This 1.2-kb locus is marked by two distinct, large insertion events: a nuclear transposition of a mitochondrial sequence and a transposition of a nonautonomous DNA transposon DNAREP1_DM. Because we could trace the evolutionary histories of these sequences, we were able to reconstruct the length evolution of this region in some detail. We sequenced this locus in all four species of the *D. melanogaster* species complex: *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana*. Although this locus is similar in size in these four species, less than 10% of the sequence from the most recent common ancestor remains in *D. melanogaster* and all of its sister species. This region appears to have increased in size through several distinct insertions in the ancestor of the *D. melanogaster* species complex and has been shrinking since the split of these lineages. In addition, we found no evidence suggesting that the size of this locus has been maintained over evolutionary time; these results are consistent with the model of a dynamic equilibrium between persistent DNA loss through small deletions and more sporadic DNA gain through less frequent but longer insertions. The apparent stability of genome size in Drosophila may belie very rapid sequence turnover at intergenic loci.

Introduction

Genomes of closely related organisms often have reasonably similar genome sizes and gene complements. Genome sizes within the *Drosophila melanogaster* species subgroup, for instance, are all roughly equivalent (Powell 1997). In this specific case, the stability of genome size is particularly intriguing, because there is a strong mutational pressure towards DNA loss at a small scale. This trend, manifest in the more frequent and longer spontaneous deletions relative to insertions, has been documented in a variety of unconstrained sequences in Drosophila (Petrov 2002*a*): "dead-on-arrival" copies of non-LTR retrotransposable elements (Petrov et al. 1998; Blumenstiel, Hartl, and Lozovsky 2002), several pseudogenes (Pritchard and Schaeffer 1997; Petrov et al. 1998; Ramos-Onsins and Aguade 1998; Robin et al. 2000), and an insertion of mitochondrial DNA into the nuclear genome (Petrov 2002*a*).

This strong bias toward DNA loss in the *D. melanogaster* lineage apparently evolved before the separation of *D. melanogaster* and *D. virilis* (Petrov, Lozovskaya, and Hartl 1996; Petrov et al. 1998), approximately 60 MYA (Russo, Takezaki, and Nei 1995). In 60 Myr, without a counterbalancing source of DNA addition, small deletions are expected to remove approximately 95% of unconstrained DNA. Thus, one possible explanation for the apparent recent stability of genome size within the *D. melanogaster* species subgroup is simply that the genomes in this subgroup are composed entirely of functional sequences.

Is it plausible that all sequences in the Drosophila genome are functional and thereby retained by purifying selection? Although this is difficult to judge, the observation that most noncoding sequences evolve very quickly (close to the expected neutral rate of evolution) casts doubt on this interpretation. For instance, a study by Bergman and Kreitman (2001) revealed the presence of only short interspersed blocks of constrained sequences within intergenic and intronic DNA. Overall, less than 30% of intergenic DNA appears to be constrained. Of course, this may be an underestimate if what matters is not the precise sequence but the presence of DNA of particular length at a particular location in the genome. Nevertheless, the current evidence suggests that a significant proportion of noncoding DNA in Drosophila is truly unconstrained at least in its exact sequence content.

Given the strong mutational pressure towards DNA loss that is surely operating in this species group, how then is this unconstrained DNA maintained? The first possibility, as alluded to earlier, is that there is purifying selection acting on the length rather than the sequence of intergenic regions. Under this model, the sequence content may often be of no selective importance, but the lengths of intergenic regions have functional significance and are, therefore, maintained by purifying selection. This model predicts that the length of an intergenic region should remain constant over evolutionary time.

The alternative explanation is that intergenic DNA in Drosophila is maintained through a dynamic equilibrium between large DNA insertions and small DNA deletions (Petrov 2002*b*). The current measurements of indel biases are limited to small (<400 bp) indels, and we know that among such indels, deletions predominate. However, if insertions are more common among large indels, they could potentially offset the loss of DNA from frequent but small deletions. Although little is known about the rate of large indels, we can surmise that insertions are likely to be more common among those large indels that reach fixation and, thus, ultimately affect the lengths of intergenic loci. In a compact, gene-rich genome, a large deletion is likely to disrupt neighboring genes (Ptak and Petrov 2002) with at

Key words: DNAREP1_DM, numt, intergenic, insertions, deletions, genome evolution.

E-mail: ndsingh@stanford.edu.

Mol. Biol. Evol. 21(4):670–680. 2004
DOI:10.1093/molbev/msh060
Advance Access publication January 22, 2004
Molecular Biology and Evolution vol. 21 no. 4

least one of its two breakpoints; these large deletions will quickly be removed by strong selection for genic maintenance. Insertions, however, only have one breakpoint, and accordingly, large insertions will have as good a probability in landing in an unconstrained place as small ones. This effect, wherein large insertions have a greater chance of reaching fixation than similarly sized deletions, should be most pronounced in short intergenic regions. Under this model of intergenic DNA maintenance, the predictions are twofold: (1) The length of an intergenic region should vary widely over evolutionary time. (2) The sequence content of orthologous intergenic loci in closely related Drosophila species could differ dramatically as a direct consequence of the balance of these two stochastic processes.

To distinguish between these two hypotheses for the maintenance of intergenic DNA and to start quantifying the rates of DNA addition through large insertions, we chose to study an intergenic region that contained potentially unconstrained DNA sequences whose evolutionary history we could trace. It is possible to identify three nuclear insertions of mitochondrial DNA (numt) into the *D. melanogaster* genome (Petrov 2002*a*); the only numt that is sufficiently long for analysis is 566 bp and was inserted on the fourth chromosome approximately 230 kb from the centromere. In addition to the numt insertion, this intergenic locus on the fourth chromosome also contains a single insertion of a nonautonomous DNA element, DNAREP1_DM.

To date, all available evidence suggests that both the numt and the copy of DNAREP1_DM are noncoding and unconstrained at the level of their sequence. Numts have never been seen to retain any coding function in metazoans (Bensasson et al. 2001), most likely because the mitochondrial genetic code in animals is distinct from the nuclear genetic code. In addition, DNAREP1_DM elements have no open reading frame and appear to have been immobile for millions of years (Kapitonov and Jurka 2003). These considerations suggested that the 1200-bp region, including both the numt and the DNAREP1_DM insertion, was a good candidate for an intergenic region that was unconstrained in terms of its sequence content.

However, the location of this locus on the fourth chromosome opened the possibility that this locus was constrained with respect to its length, because it is quite close to genes on either side. The annotated genes that are nearest to this locus are *Crk* (∼500 bp downstream) and CG31998 (∼3 kb upstream). *Crk* (synonym: CG1587) seems to be an SH3/SH2 adaptor protein involved in signal transduction and is expressed in the embryo. Although CG31998 (synonyms: CG11578 and CG11572) has no known function to date, the gene prediction is supported by EST data. Interestingly, these two genes are coded in opposite directions, and, accordingly, the region between them, in which our locus is found, is in the potential 5′ upstream regulatory region for both genes. As a result, although this 1200-bp region may not be under selection for its sequence content, its length may in fact be constrained.

However, our results failed to show any evidence of selective maintenance of the ancestral length of this region. To the contrary, we argue that this region expanded in the ancestor of the *D. melanogaster* species complex through the insertions of the numt and several copies of DNAREP1_DM and has been going through persistent, and apparently random, shrinkage since then. Our results demonstrate the power of persistent DNA loss and support the predictions of the model of a dynamic equilibrium between rare but large insertions and more common but smaller deletions. The apparent stability of genomes in the *D. melanogaster* species complex belies a very rapid sequence turnover; although the amount of intergenic DNA may often be similar in sister species, very little of it may prove to be truly orthologous, even in very closely related species.

Materials and Methods

Fly Strains and Genomic DNA Extraction

Seventeen strains of *D. melanogaster* were used in this study: five from Ann Arbor, Mich. (A1, A3, A6, A8, and A18) (gift from G. Gibson); seven from Davis, Calif. (WI1, WI15, WI41, WI45, WI68, WI83, and WI69) (gift from S. Nuzhdin); and five representing worldwide samples (W2 [Bermuda], W7 [New York], W9 [Australia], W22 [Georgia, USA], W31 [Nairobi]) (gift from G. Gibson). We extracted genomic DNA from single males taken from these 17 strains of *D. melanogaster*, and one strain each of *D. simulans*, *D. mauritiana*, and *D. sechellia* (Arizona Stock Center species 14021-0248.3) according to protocol described by Greg Gloor and William Engels (personal communication). Each fly was crushed with the end of a pipette tip and subsequently immersed in a buffered solution (10 mM Tris-Cl pH 8.2, 1 mM EDTA, 25 mM NaCl, 200 μg/ml proteinase K). This was incubated at 37°C for 30 min and then at 95°C for 2 min to inactivate the proteinase K.

PCR and Sequencing

Amplifying conditions for each of the four species are as follows: *D. melanogaster*, *D. simulans*, and *D. mauritiana*, 94°C for 2 min, 30 cycles of 94°C for 30 s, 59°C for 30 s, 72°C for 30 s, and a final extension of 72°C for 7 min; *D. sechellia*, 94°C for 2 min followed by 37 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 2 min, followed by a final extension of 72°C for 7 min. All PCR reactions were 20 μl, and each contained 2 μl 10X Quiagen PCR Buffer, 2 μl 1.25 mM dNTP, 0.2 μl of each 20 μM primer, 0.2 μl Quiagen Taq, 13.4 μl H_2O, and 2 μl genomic DNA. Amplifying primers (3844mt ±) were designed from the sequence of *D. melanogaster* obtained from GenBank, and internal primers were eventually designed from our own sequence data. Two internal primers were designed for *D. melanogaster* (3844IntF/R), one was designed for *D. simulans* (3844SimF), and two were designed for *D. sechellia* (3844SechF1/R1). Primer sequences, 5′ to 3′ are as follows: 3844mt+, CGA ATA AGC CAA GAA CCC TAA; 3844mt−, CTC CGG TCG CTA TCT GAT; 3844IntF, AAT TGGT TAA AAC TTA ACG AAA AT; 3844IntR, TCT TGT AAA TTT CTA TCG ATT TG, 3844SimF, CTC GAC GTT CAT ACG

GAC; 3844SechF1, TAT TTT ATA TGT AAA AAT TGC, 3844SechR1, AGA GAT TTA CTA GAT TCG TTG. PCR reactions were enzymatically cleaned with exonuclease I and shrimp alkaline phosphatase, and were cycle-sequenced in half-strength half-reactions with Big Dye under standard cycling conditions. These reactions were precipitated using ethanol and $MgSO_4$ and sequenced on an ABI 377 sequencer.

DNAREP1_DM Analysis

To test hypotheses regarding the evolutionary history of DNAREP1_DM, we implemented a bioinformatic approach. We used NCBI's version of BlastN, blasting the reported consensus sequence for DNAREP1_DM (Kapitonov and Jurka 1999) against the *D. melanogaster* genome with the following parameters. The reward for a match, penalty for a mismatch, gap-opening penalty, and gap-extension penalty were 5, -5, 10, and 2, respectively. In addition, we used a word size of 23 bp, a Blast extension dropoff of 15, a final dropoff of 10, and an e-value of 0.01. We retrieved 5,000 one-line descriptions and 5,000 alignments for our genome-wide analyses of this element. Pairwise distances among elements were calculated based on the alignments performed in BlastN, discounting insertions and deletions. The parameters implemented above restrict retrievals to sequences differing by no more than 30% from the sequence query; in this respect, our search criteria were conservative.

To establish relationships of orthology and paralogy among the copies of DNAREP1_DM at our locus, we retrieved 76 fragments of DNAREP1_DM from the fourth chromosome of the *D. melanogaster* genome using all of the default parameters in NCBI's BlastN. Because these parameters are highly restrictive, the distribution of pairwise distances among these copies is extremely conservative for our purposes.

Sequence Alignment and Statistical Analyses

Sequences were aligned using a combination of Sequencher version 3.1.1 and MacVector version 7. Sequences were considered properly aligned if there was identity of at least 80% over a stretch of nucleotides. Sequences were screened for repetitive elements using RepeatMasker (Smit, A.F.A. & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html), which led to the identification of DNAREP1_DM elements in our sequence data. MEGA version 2.1 (Kumar, Tamura, and Nei 1994) was used to calculate π and Tajima's *D* statistic (Tajima 1989). PAUP version 4.0b9 was used for reconstruction of the mitochondrial phylogeny, and was also used to compute pairwise distances among the 76 fourth chromosome copies of DNAREP1_DM.

The tests of goodness of fit were conducted using G-tests (Sokal and Rohlf 1997). Where necessary, the expectations from continuous distributions were converted into expectations for integer counts. For comparisons of molecular rates of evolution, this transformation involved calculating the expected number of substitutions over a particular amount of time for a sequence and comparing it to the observed number of substitutions. In addition, one goodness-of-fit test was performed on levels of polymorphism ($\Theta = 4N_e\mu$); this test compared the number of segregating sites in sequences (whose lengths were known). The fit of certain data to the Poisson distribution was ascertained by testing whether the ratio of the observed variance to the observed mean was significantly different from one. The significance was derived from the χ^2 distribution with $n-1$ degrees of freedom, where n is the number of observations.

Results

Sequence of the Chosen Intergenic Region in the *D. melanogaster* Species Complex

We attempted to amplify the 1.2-kb intergenic locus on the fourth chromosome in several members of the *Drosophila melanogaster* species subgroup: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, and *D. yakuba*. We were unable to amplify this locus in *D. yakuba*, presumably because the primer sites are no longer recognizable, but the sizes of the amplified fragments in the remaining four species were within the same order of magnitude, ranging from 0.95 kb in *D. mauritiana* to 1.8 kb in *D. simulans*. We sequenced the amplified fragments in each of these four species.

Our sequence data revealed evidence of high sequence turnover because only 169 bp of the reconstructed ancestral sequence had been retained in all four species (fig. 1). We inferred that the minimum size of this locus in the most recent common ancestor (MRCA) of these four species was 1,847 bp. This estimate is highly conservative; in all likelihood this locus was over 2.3 kb in the MRCA. Because of the lack of sequence similarity, we were initially concerned that we were not amplifying orthologous loci in these four species. However, the PCR reactions were highly specific, with consistent production of one discrete band of approximately the correct size in each of these four species. In addition, the easily alignable sequences are located immediately adjacent to the primer sites on each side. Finally, comparing the sequence data from *D. melanogaster* and *D. simulans* in the region of overlap yielded a Jukes-Cantor substitution distance of 0.158, which is entirely consistent with interspecific divergence calculated from other pseudogene loci (table 1). Below we describe how we inferred the history of this region in detail.

numt Analysis

The sequence of the studied region in *D. melanogaster* contains a 566-bp insertion of mitochondrial DNA that is absent in the orthologous region of each of the other three studied species (fig. 1). To determine the timing of the numt insertion, we reconstructed its phylogenetic history relative to the homologous mitochondrial region in the *D. melanogaster* species group. We used both parsimony and maximum-likelihood (HKY85) criteria. Exhaustive searches using either criterion resulted in the same unique best tree (fig. 2), which shows that the numt inserted in the ancestor of *D. melanogaster* species

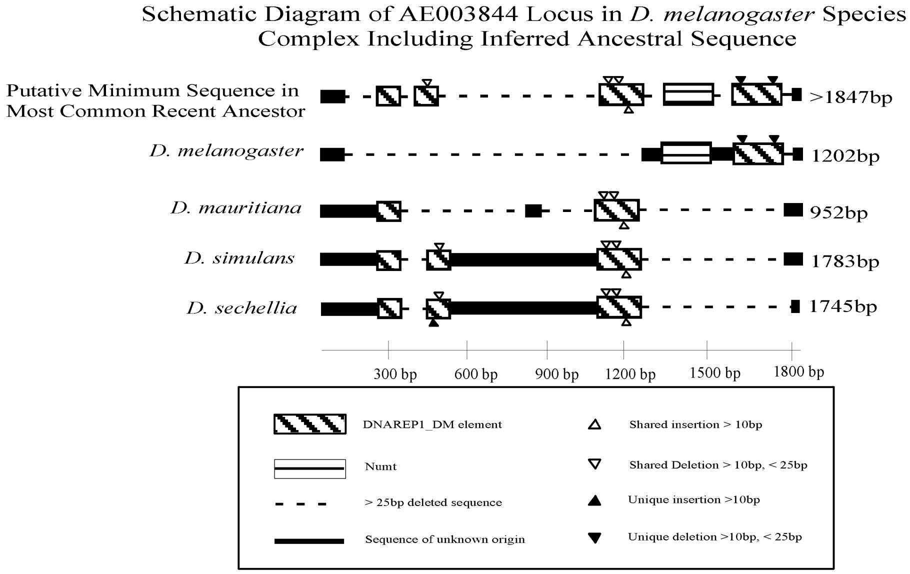

Schematic Diagram of AE003844 Locus in D. melanogaster Species Complex Including Inferred Ancestral Sequence
Putative Minimum Sequence in Most Common Recent Ancestor
>1847bp
D. melanogaster
1202bp
D. mauritiana
952bp
D. simulans
1783bp
D. sechellia
1745bp
300 bp
600 bp
900 bp
1200 bp
1500 bp
1800 bp
DNAREP1_DM element
Numt
> 25bp deleted sequence
Sequence of unknown origin
Shared insertion > 10bp
Shared Deletion > 10bp, < 25bp
Unique insertion >10bp
Unique deletion >10bp, < 25bp

FIG. 1.—Schematic representation of AE003844 locus in the four members of the *D. melanogaster* species complex, as well as the inferred ancestral sequence. The nuclear transposition of a mitochondrial sequence (numt) is 566 bp, and the first, second, and third copies of DNAREP1_DM are 340 bp, 422 bp, and 350 bp, respectively. The sequence of the putative ancestor, however, is the minimum estimate, based on remaining sequence in the four descendant species presented.

complex after its branching from the *D. yakuba* lineage (~6 MYA) and before the branching of the *D. simulans* lineage (~2.3 MYA) (Russo, Takezaki, and Nei 1995).

To estimate the confidence in the timing of the numt insertion, we went down the list of the maximum-likelihood trees until we found one (the third down the list) that showed the numt inserting in the *D. melanogaster* lineage since its split from *D. simulans*. The Kishino-Hasegawa test (Hasegawa and Kishino 1990) failed to show that the best tree was significantly better than this tree ($P = 0.1$).

However, in addition to the phylogenetic information implicit in the sequence of the numt, we can also investigate whether the lengths of the numt branches in the two trees are consistent with the known rates of molecular evolution of unconstrained Drosophila sequences. The hypothesis implied by the best tree suggests that the numt is 4.2 ± 0.95 Myr old, inserting in the nuclear genome approximately halfway between 6.2 and 2.3 MYA. Given that the numt is 11.4% different from its mitochondrial ancestor, we can then estimate the rate of evolution in the numt to be between 22×10^{-3} and 35×10^{-3} substitutions/site/Myr. This rate is similar to the rate reported for other Drosophila pseudogenes (33.3×10^{-3} substitutions/site/Myr) (table 1).

In contrast, the insertion of the numt after the split of the *D. melanogaster* and *D. simulans* lineages implies that the numt is less than 2.3 Myr old and, thus, has been evolving very quickly (faster than 50×10^{-3} substitutions/site/Myr). This rate is significantly higher than the average rate reported for other Drosophila pseudogenes ($P = 0.02$, G-test). It is also significantly higher that the rate of evolution at the shared sequence at this locus ($P < 0.001$, G-test), suggesting that the rate of evolution in this region is not generally elevated. On balance, these results strongly favor the hypothesis of the numt inserting approximately 4 MYA in the ancestor of the *D. melanogaster* species complex. The absence of the numt in *D. simulans*, *D. mauritiana*, and *D. sechellia*, therefore, implies the loss of this sequence in those lineages since their split from the *D. melanogaster* lineage.

DNAREP1_DM Analysis

DNAREP1_DM was originally described by Kapitonov and Jurka (1999) as a 594-bp nonautonomous DNA transposon. This element is ubiquitous in the *D. melanogaster* genome; previous analysis suggested that there were several thousand copies in the genome (Kapitonov and

Table 1
Divergence Between *D. melanogaster* and *D. simulans* at Pseudogene Loci

Locus	Number of Sites[a]	Divergence[b]	Estimated Substitution Rate[c]	Citation
CecΨ1	368	0.16	34.8×10^{-3}	Ramos-Onsins and Aguadé 1998
CecΨ2	401	0.15	32.6×10^{-3}	Ramos-Onsins and Aguadé 1998
LcpΨ	356	0.149	32.4×10^{-3}	Pritchard and Schaeffer 1997
AE003844	169	0.157	34.1×10^{-3}	

[a] Number of nucleotides included in the divergence calculation.

[b] Jukes-Cantor corrected distance (substitutions per site) between *D. melanogaster* and *D. simulans*.

[c] Estimated substitution rate per site per Myr.

Jurka 2003), and our own analysis, which was restricted to mostly euchromatic sequence with conservative search parameters, yielded almost 1,100 copies of this element scattered all over the genome. The sequence of our region in *D. melanogaster* contains a single copy of DNAREP1_DM, whereas the sequences in *D. simulans*, *D. mauritiana*, and *D. sechellia* each contain two copies of DNAREP1_DM. To determine the history of the acquisition and loss of these elements in our region, we needed to answer several questions. For instance, we needed to determine both when the DNAREP1_DM elements entered the genome and whether DNAREP1_DM has been transpositionally active in the recent past. In addition, we needed to establish the paralogy and orthology relationships among the seven identified copies of DNAREP1_DM in our region.

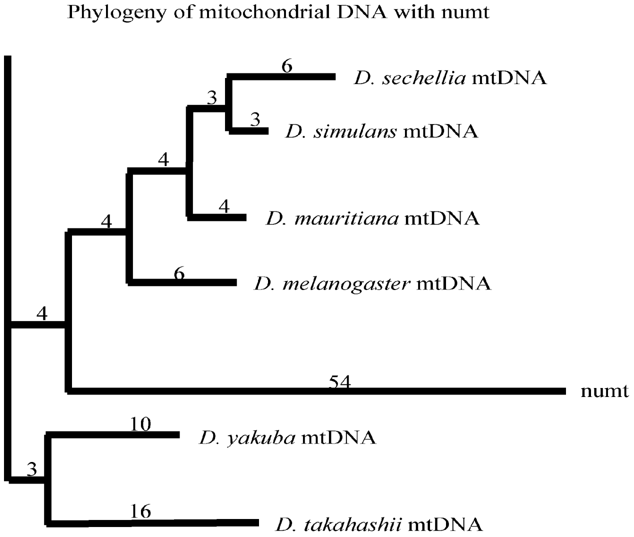

Phylogeny of mitochondrial DNA with numt

FIG. 2.—Phylogenetic reconstruction using mitochondrial DNA, with *D. erecta* (not shown) as an outgroup. This is the unique best tree using parsimony and maximum-likelihood criteria. Branch lengths are proportional to the number of changes mapping to each branch.

Evidence for the Burst of DNAREP1_DM Transposition in the Ancestor of the *D. melanogaster* Species Complex

Initial analysis suggested that DNAREP1_DM was mobilized in a burst of transposition in a common ancestor of the Drosophiloidae and has remained inactive ever since (Kapitonov and Jurka 2003). This model of a single burst of transposition followed by independent neutral evolution of each copy predicts a starlike phylogeny and a Poisson distribution of pairwise distances between extant DNAREP1_DM copies and their ancestor. To test this prediction, we retrieved 1,087 distinct copies of DNAREP1_DM from the sequenced genome of *D. melanogaster* and computed pairwise distances between each copy and the reported consensus (and presumably ancestral) sequence. Consistent with the hypothesis of a single burst of transposition some time in the past, this distribution has a single peak (at 15.2% divergence) (fig. 3). The distribution is different from the Poisson—the average divergence (15.2%) is significantly smaller than the variance of the distribution (29.16%) ($P < 0.05$, χ^2 distribution). However, the biggest difference is in the right hand tail, which may be an artifact of our methodology. It is clear that our search for DNAREP1_DM copies using Blast against the sequenced *D. melanogaster* genome biases us toward finding the least-diverged copies, and, given our search parameters, we were unable to retrieve copies diverged by more than 30%. As a result, we are likely to miss many of the more-diverged copies and, therefore, underestimate the true age of the DNAREP1_DM family. Note, however, that because we argue that the DNAREP1_DM copies inserted into the studied locus before the separation of the species in the *D. melanogaster* complex, this sampling bias toward younger copies of DNAREP1_DM makes our analyses conservative.

Importantly, we would not expect this general shape under a model of continual transposition, which should instead produce an overabundance of recent (least-diverged) elements because of the combined effects of mutation accumulation and ascertainment bias. If we take the average divergence of 15.2% to correspond to the main burst of DNAREP1_DM activity, we can infer that most of the copies of DNAREP1_DM inserted approximately 4.6 MYA, well before the estimated time of divergence of the *D. melanogaster* and *D. simulans* lineages (2.3 ± 0.65 MYA). The hypothesis that the average divergence of DNAREP1_DM is equal to or less than the divergence expected for a pseudogene (table 1) inserted at the time of the split of the *D. melanogaster* and *D. simulans* lineages can be rejected with high confidence ($P < 0.001$, G-test).

It is entirely possible, however, that DNAREP1_DM has been active, albeit at a lower level, more recently than 4.6 MYA. Assuming that the active sequence has remained the same, we can estimate the proportion of recent transpositions by looking for DNAREP1_DM elements that are more similar to the consensus than expected under the Poisson distribution. The expected divergence of orthologous pseudogenes in *D. melanogaster* and *D. simulans* is 7.6%. Our analysis suggests that no more than 4% of all of the copies of DNAREP1_DM in the *D. melanogaster* genome could have transposed since the speciation of the *D. melanogaster* complex.

It is similarly possible that some old DNAREP1_DM elements have been duplicated (or even transposed) since their original transposition. Such copies might look old in a comparison with the ancestor, yet would have been inserted in our region recently. We can quantify the likelihood of this possibility by comparing the number of elements that are more similar to each other than expected under the Poisson distribution. We used the sequence of the copy of DNAREP1_DM present at our locus in *D. melanogaster* as a query for a Blast search and retrieved the 245 best hits for this sequence. This distribution of pairwise distances, with mean 19.9% and variance 17.9%, is not significantly different from a Poisson ($P > 0.9$, χ^2 distribution). If we use our sample mean to generate a Poisson distribution, the expected number of comparisons yielding pairwise divergences below and above 7.6% are 0.2 and 244.8, respectively, whereas we observed, respectively, 0 and 245 such comparisons. This strongly suggests that there have not been many recent duplications of this element. Taken together, these analyses indicate

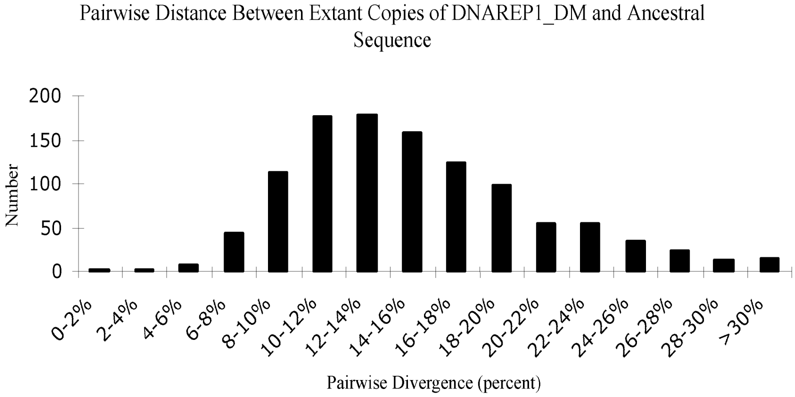


FIG. 3.—Distribution of Jukes-Cantor corrected pairwise distances between 1,087 copies of DNAREP1_DM and the ancestral sequence. This distribution has a mean of 15.2% and a variance of 29.2%.

that all observed copies of DNAREP1_DM at the studied locus inserted in an ancestor of the *D. melanogaster* species complex.

Paralogy and Orthology Relationships Among the Seven Identified Copies of DNAREP1_DM

Based on sequence similarity within the repeat as well as flanking sequence around each copy of the repeat, we tentatively determined that there were three distinct copies of DNAREP1_DM present, with two copies each in *D. simulans*, *D. sechellia*, and *D. mauritiana*, and a third copy in *D. melanogaster* (fig. 1). To confirm these assignments of orthology and paralogy, we developed a more rigorous metric based on pairwise distances among paralogous copies of DNAREP1_DM. To ensure that our inferences were as precise as possible, we based our metric for distinguishing between orthology and paralogy solely on copies of DNAREP1_DM on the fourth chromosome of the *D. melanogaster* genome. We retrieved 76 fragments of DNAREP1_DM from the fourth chromosome, each of which were at least 50 bp in length and were taken from the 5′ end of the repetitive element. Some fragments did not overlap sufficiently for analysis and have been excluded from the distribution. The distribution of the uncorrected pairwise distances among these sequences (fig. 4) was used to generate expectation under paralogy. We used the most restrictive Blast search parameters (see *Materials and Methods*) and did not correct for multiple hits to bias our estimate toward a higher proportion of similar copies of DNAREP1_DM. Such a bias is conservative for our purposes, given that we argue that all DNAREP1_DM copies in our locus inserted before the speciation of the *D. melanogaster* species complex.

Pairwise differences (table 2) between the putatively orthologous versions of the first and the second copies of the element in *D. simulans*, *D. sechellia*, and *D. mauritiana* range from approximately 2% to 6%. These distances are consistent with the expected divergence between orthologous *D. simulans* and *D. mauritiana* sequences, (5.9%, given a divergence time of 0.9 Myr and the pseudogene substitution rate of 33.3×10^{-3} substitution/site/Myr [see table 1]). However, they are inconsistent with the expected divergences among the paralogous copies, given that fewer than 3% of comparisons differ by 6% or fewer (fig. 4). We are thus confident that each of these two copies of the element is orthologous in *D. simulans*, *D. sechellia*, and *D. mauritiana*.

In contrast, the presumptively paralogous copies differ from approximately 27% to 37% from one another. These distances are consistent with the expectations of paralogy. Each of these copies is approximately 15% divergent from the ancestor (results not shown), and, thus, the pairwise distance is expected to be around 30%. Additionally, the pairwise distance between the copy of DNAREP1_DM in *D. melanogaster* and either of the copies in the other species (table 2) is inconsistent with orthology, given that this distance is significantly greater than expected for orthologous pseudogenes (table 1) in these species ($P = 0.002$, G-test), as well as being notably higher than the 16% divergence of the only region we could align between *D. melanogaster* and *D. simulans*.

Although it is possible that different pseudogenes evolve at different rates, there is no significant difference in the substitution rate among the known pseudogenes ($P > 0.75$ for all pairwise comparisons, G-test). Moreover, the fact that the divergence among the orthologous copies of DNAREP1_DM in *D. simulans*, *D. sechellia*, and *D. mauritiana* is entirely consistent with expectation argues against DNAREP1_DM in general evolving at a higher rate than other pseudogenes. Altogether, our analyses suggest that the copy of DNAREP1_DM in *D. melanogaster* is indeed distinct from both copies present in its sister species, and as a result, that there were at least three copies of DNAREP1_DM at the studied locus in the MRCA of the *D. melanogaster* species complex.

Estimation of the Deletion/Insertion Biases

We do not know the original length of the numt insertion and, thus, cannot evaluate the rate of deletions at the edges of the original numt. There are six recognizable internal deletions (three of 1 bp and one each of 13 bp, 14 bp, and 27 bp) and no internal insertions in the numt. The relative rates of point substitutions to deletions (11.5:1) and deletions to insertions (6:0) are not signifi-

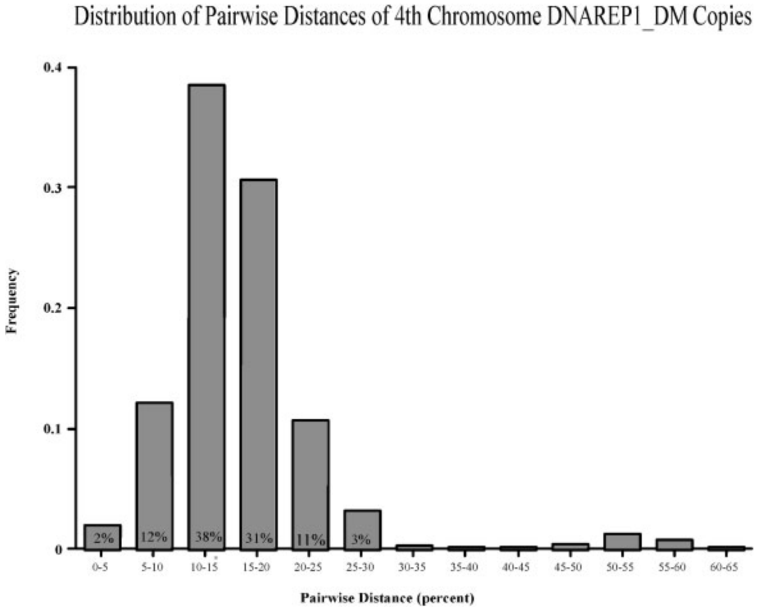


FIG. 4.—Distribution of pairwise distances among 76 copies of DNAREP1_DM on the fourth chromosome. Comparisons without sufficient overlap were excluded. This distribution has a mean divergence of 15.9%, with a standard deviation of 7.98%, and reflects a conservative estimate of these distances. Not only are the distances uncorrected but the entire distribution is shifted toward higher similarity because of ascertainment bias. Despite this, note that fewer than 3% of comparisons yield divergence of 6% or less.

cantly different from those ratios reported for pseudogenes ($P = 0.98$, G-test) (Pritchard and Schaeffer 1997; Petrov et al. 1998; Ramos-Onsins and Aguade 1998; Robin et al. 2000) or transposable elements ($P = 0.80$, G-test) (Petrov, Lozovskaya, and Hartl 1996; Blumenstiel, Hartl, and Lozovsky 2002). The distribution of the lengths of the deletions is also very similar to that observed for other pseudogenes (Petrov et al. 1998), with half of the deletions smaller than 10 bp and half of the deletions longer than 10 bp.

With regard to DNAREP1_DM, we are less confident in the exact sequence of the ancestor, but we do have confidence in its length. We, therefore, decided to avoid the identification of individual deletion and insertion events and instead compare the remaining lengths of DNAREP1_DM elements with the predicted length under exponential deletion-induced decay. The length of a given element is expected to contract exponentially according to the formula: (starting length) × e^{-dt}, where d is the rate of DNA loss per substitution per bp (which is conservatively estimated at 3.8 bp per substitution per bp [Blumenstiel, Hartl, and Lozovsky 2002] in *D. melanogaster*) and t is time measured in point substitutions per bp. Because DNAREP1_DM elements are on average 15.2% divergent from the ancestral sequence, they should be approximately 61% in length now ($e^{[-3.8 \times .152]} = 0.61$). The lengths of the elements, varying from 57% to 76%, are roughly consistent with these predictions.

Additionally, we can identify individual indels in a more shallow comparison, between *D. simulans* and *D. sechellia* (divergence time approximately 0.9 Myr). In this comparison, our observations are also consistent with other studies. The relative ratios of nucleotide substitutions to deletions (19:2) and deletions to insertions (2:0) are similar to ratios from both pseudogene studies ($P = 0.98$ and $P = 0.80$, G-test for substitutions versus deletions and deletions versus insertions, respectively) and transposable elements ($P = 0.52$, $P = 0.80$, G-test for substitutions versus deletions and deletions versus insertions, respectively).

Table 2
Jukes-Cantor Pairwise Distances[a] Among DNAREP1_DM Copies

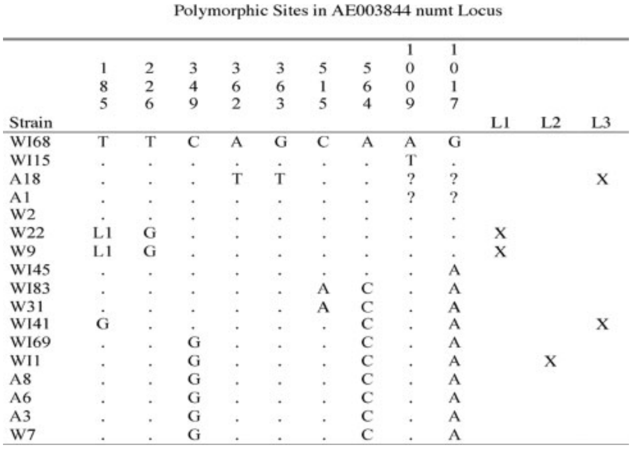
Sim Copy 1[b]	Sech Copy 1[b]	Maur Copy 1[b]	Sim Copy 2[c]	Sech Copy 2[c]	Maur Copy 2[c]	Mel Copy[d]
Sim Copy 1						
Sech Copy 1	**0.046**					
Maur Copy 1	**0.054**	**0.056**				
Sim Copy 2	0.327	0.323	0.357			
Sech Copy 2	0.350	0.367	0.365	**0.049**		
Maur Copy 2	0.307	0.303	0.330	**0.020**	**0.037**	
Mel Copy	0.325	0.330	0.358	0.283	0.321	0.270

[a] Jukes-Cantor corrected distance (substitutions per site) among seven copies of DNAREP1_DM at this locus. Entries in bold are comparisons yielding divergence estimates less than or equal to 6%.

[b] First copy of DNAREP1_DM at this locus in *D. simulans*, *D. sechellia*, and *D. mauritiana*, respectively.

[c] Second copy DNAREP1_DM at this locus in *D. simulans*, *D. sechellia*, and *D. mauritiana*, respectively.

[d] Sole copy of DNAREP1_DM at this locus in *D. melanogaster*.



Polymorphic Sites in AE003844 numt Locus

								1	1			
	1	2	3	3	3	5	5	0	0			
	8	2	4	6	6	1	6	0	1			
	5	6	9	2	3	5	4	9	7			
Strain										L1	L2	L3
WI68	T	T	C	A	G	C	A	A	G			
WI15	.	.	.	.	.	.	.	T	.			
A18	.	.	.	T	T	.	.	?	?			X
A1	.	.	.	.	.	.	.	?	?			
W2	.	.	.	.	.	.	.	.	.			
W22	L1	G	.	.	.	.	.	.	.	X		
W9	L1	G	.	.	.	.	.	.	.	X		
WI45	.	.	.	.	.	.	.	.	A			
WI83	.	.	.	.	.	A	C	.	A			
W31	.	.	.	.	.	A	C	.	A			
WI41	G	.	.	.	.	.	C	.	A			X
WI69	.	.	G	.	.	.	C	.	A			
WI1	.	.	G	.	.	.	C	.	A		X	
A8	.	.	G	.	.	.	C	.	A			
A6	.	.	G	.	.	.	C	.	A			
A3	.	.	G	.	.	.	C	.	A			
W7	.	.	G	.	.	.	C	.	A			

FIG. 5.—Polymorphic sites within the 17 strains of *D. melanogaster* sequence. Three length polymorphisms are denoted L1, L2, and L3; missing data are represented by a question mark (?). Strains W2, W7, W9, W22, and W31 represent worldwide samples of *D. melanogaster*, strains A1, A3, A6, A8, A18 are from Ann Arbor, Michigan, and WI1, WI15, WI41, WI45, WI68, WI83, WI69 are strains from Davis, California. A total of nine haplotypes were detected based on these sequence polymorphisms.

Polymorphism Within *Drosophila melanogaster*

We also decided to compare patterns of polymorphism in our region with other known fourth chromosome loci. Using our amplifying primers, approximately 250 bp upstream of the numt and 400 bp downstream, we sequenced the resulting 1.2-kb product in 17 *D. melanogaster* strains. Fourteen strains came from North America: five strains from Ann Arbor, Mich. (A1, A3, A6, A8, and A18), seven strains from Davis, Calif. (WI1, WI15, WI41, WI45, WI68, WI83, and WI69), one strain from New York (W7), and one strain from Georgia, USA (W22). The remaining strains were collected in Australia (W9), Bermuda (W2), and Kenya (W31).

The polymorphic sites are shown in figure 5. Three sequence-length polymorphisms were detected at this locus, one of 1 bp, one of 6 bp, and one of 156 bp, which was associated with a T to G transversion at the nucleotide position immediately 5′ of the deleted bases. Because these length polymorphisms were in regions of this locus that flank the numt and DNAREP1_DM, we cannot determine whether they are products of insertions or deletions. We also detected nine segregating single-nucleotide polymorphisms. The segregating sites (including indel polymorphisms) fall into nine distinct haplotypes, with no clear pattern of geographical structure. There are no recombination events that can be detected among the haplotypes. The number of observed haplotypes is within the range expected under neutrality, given the observed number of segregating sites; both the number of haplotypes (K) and the haplotype diversity (H) are within the 95% confidence interval for the expectation of these parameters under the assumption of no recombination (Depaulis and Veuille 1998). Although sequences for two haplotypes have missing data, the neutrality of the haplotype data is robust to possible findings of any of the polymorphic states within the missing data.

The estimates of Θ ($4N_e\mu$) are 0.0027 per nucleotide based on S (the number of segregating sites) and 0.0026 based on π (the average pairwise difference). These estimates are not significantly different from each other and, therefore, are consistent with neutrality (Tajima's D statistic of -0.158; $P > 0.1$) (Tajima 1989). Calculating S and π using solely numt sequence, solely flanking sequence, or the entire region does not yield significantly different estimates, and this level of nucleotide polymorphism is similar to that estimated from other regions on the fourth chromosome (Wang et al. 2002).

Discussion

History of the Region

Our analysis suggests that the sequence of the studied locus in the ancestor of the *D. melanogaster* species complex had at the minimum a single numt (566 bp), three copies of DNAREP1_DM (340 bp, 422 bp, and 350 bp, respectively), and at least 169 bp of additional sequence (the only alignable sequence that remains in all of the species) (fig. 1). Thus, the total length of this region in the inferred MRCA (2.3 MYA) was at a minimum 1,847 bp and most likely closer to 2,329 bp, given the sequence shared by several species in this complex. Based on this putative ancestral sequence, we can infer that during the course of evolution, the locus suffered one large deletion (or a number of small deletions) in the 5′ end of the inferred ancestral sequence in *D. melanogaster*, removing the first two copies of DNAREP1_DM, as well as the sequence between these two copies. In addition, there must have been another large deletion or several small deletions in the 3′ end of the sequence in the ancestor of the remaining species, removing both the numt and the third copy of DNAREP1_DM (fig. 6). The overall pattern towards DNA loss is reflected in each of the four species of the *D. melanogaster* species complex; of the 2,329 bp (or 1,847 bp) in the ancestral sequence, 1,202 bp (52% to 65%) remain in *D. melanogaster*, 952 bp (41% to 52%) remain in *D. mauritiana*, 1,783 bp (77% to 97%) remain in *D. simulans*, and 1,745 bp (75% to 94%) remain in *D. sechellia*. Remarkably, of the remaining sequence, only 169 bp (7% to 9%) has been retained from the most recent common ancestor by all four species. It is important to note that our inference of the length of the ancestral sequence is almost certainly an underestimate, suggesting that an even smaller proportion of orthologous sequence has been retained in these species.

Tempo and Mode of Molecular Evolution at the Studied Intergenic Region

Pseudogenes and transposable elements have been used as tools to study neutral patterns of substitution in Drosophila (Pritchard and Schaeffer 1997; Petrov et al. 1998; Ramos-Onsins and Aguade 1998). These studies revealed that the relative rates of nucleotide substitutions, deletions and insertions, as well as the sizes of indels, are relatively constant across unconstrained loci (Petrov 2002*a*). Because these estimates in Drosophila have proved robust, we analyzed the rates and patterns of sub-

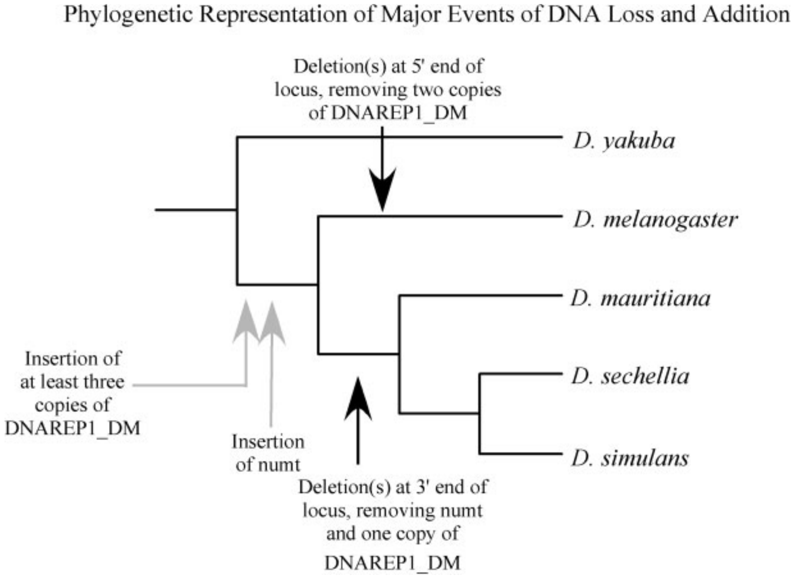

Phylogenetic Representation of Major Events of DNA Loss and Addition
Deletion(s) at 5' end of locus, removing two copies of DNAREP1_DM
D. yakuba
D. melanogaster
D. mauritiana
D. sechellia
D. simulans
Insertion of at least three copies of DNAREP1_DM
Insertion of numt
Deletion(s) at 3' end of locus, removing numt and one copy of DNAREP1_DM

FIG. 6.—Species tree of the *D. melanogaster* species subgroup, based on mitochondrial DNA. Significant DNA loss and addition events are denoted by arrows, with black arrows representing major events of DNA loss and gray arrows depicting major DNA acquisitions. Both the numt and at least three copies of DNAREP1_DM were present in the common ancestor of *D. melanogaster* and its three sister species, *D. simulans*, *D. sechellia*, and *D. mauritiana*.

stitutions at our locus to determine whether they deviate significantly from those estimated in other noncoding sequences in Drosophila. Significant acceleration of divergence or length evolution would implicate positive selection or an elevation of the mutation rate. Conversely, a reduction in rates of molecular evolution would suggest the involvement of purifying selection or a reduction in the rates of mutation.

Interestingly, all of the patterns of substitution at our locus matched the expectations based on other unconstrained Drosophila sequences. For instance, we estimated the rate of divergence in the numt to be 27.2×10^{-3} substitutions/site/Myr, which is very close to the rate found in other pseudogenes (table 1). A very similar rate of evolution was observed for orthologous copies of DNAREP1_DM. Additionally, the length evolution of both the numt and the extant copies of DNAREP1_DM also conformed to the expectation derived from neutral sequences in Drosophila; indel sizes and rates matched the expectations deduced from several studies of noncoding DNA in this species group. Furthermore, the overall impact of deletions on the size of DNAREP1_DM copies since their insertion approximately 4.6 MYA is also in agreement with predictions based on previous work. As a result, we see no evidence of either changed mutational patterns in our locus or of selection substantially changing the rates or patterns of molecular evolution.

The Pattern of Intraspecific Variation at the Studied Locus

There has been a substantial effort devoted to understanding the dynamics of nucleotide polymorphism within *D. melanogaster*, particularly on the fourth chromosome, which facilitates comparative analysis to ensure that our region did not possess exceptional levels of nucleotide polymorphism. Whereas early studies revealed the absence of any observable genetic variation in the very proximal end of the fourth chromosome (Berry, Ajioka, and Kreitman 1991; Hilton, Kliman, and Hey 1994), a recent study (Wang et al. 2002) revealed that the fourth chromosome is organized into alternating blocks of high and low polymorphism (fig. 7). Although Wang et al. (2002) sequenced 18 gene regions, most of their efforts were concentrated in the half of the chromosome distal to the centromere, whereas our region is located in the proximal half of the chromosome. The two loci already characterized at the population level with respect to nucleotide polymorphism that are closest to our region are CG1710 (Wang et al. 2002) and *ankyrin* (Jensen, Charlesworth, and Kreitman 2002), both located at least 100 kb from the studied region. *Ankyrin*, which is located 140 kb 3′ of our region showed extremely low levels of variation, whereas CG1710 (located 100 kb 5′) showed relatively high levels. Our region showed an amount of variation statistically indistinguishable from that in CG1710 (estimates of Θ are 0.0019 and 0.0027 per nucleotide for CG1710 and our region, respectively; $P = 0.9$, G-test) suggesting that the proximal half of the fourth chromosome, too, may be organized into blocks of high and low variation. Most relevant to this study, however, the amount of genetic variation at the studied region does not appear to be exceptional.

Maintenance of Intergenic DNA in Drosophila

Although the rapid loss of DNA through small deletions in Drosophila implies that unconstrained DNA should be quickly eliminated from Drosophila genomes, this does not seem to be the case. The *D. melanogaster* genome possesses substantial amounts of apparently unconstrained DNA. Estimates from the whole genome

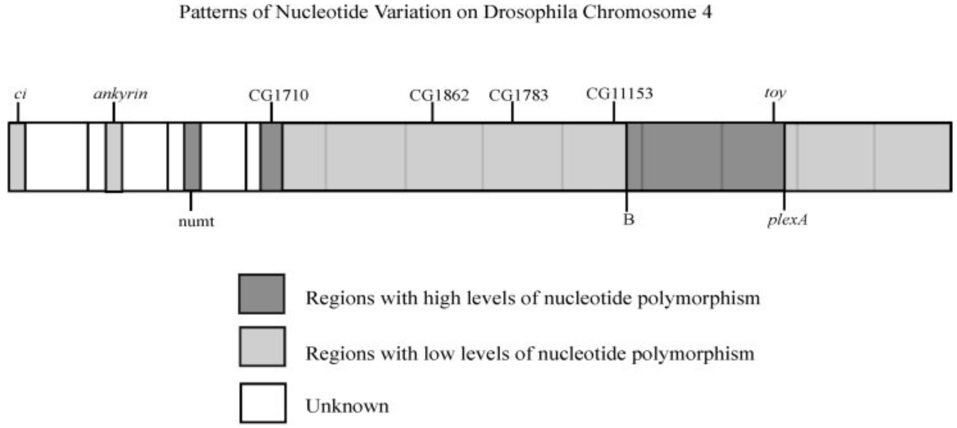

Patterns of Nucleotide Variation on Drosophila Chromosome 4
ci
ankyrin
CG1710
CG1862
CG1783
CG11153
toy
numt
B
plexA
Regions with high levels of nucleotide polymorphism
Regions with low levels of nucleotide polymorphism
Unknown

FIG. 7.—Representation of the patterns of variation on the fourth chromosome of the Drosophila genome, based on Wang et al. (2002) and Jensen, Charlesworth, and Kreitman (2002). The region of the chromosome distal to the centromere has a large region with markedly high levels of variation, and is flanked by regions of markedly low variation. This locus is roughly drawn to scale, with lines marking each 100 kb of sequence.

sequence suggest that more than 70% of the Drosophila genome is nongenic DNA (Adams et al. 2000). The observations that genomes in other Drosophila species are similar in size and rarely smaller than the *D. melanogaster* genome further suggest that they also have similar amounts of intergenic DNA. How is intergenic DNA maintained in the face of such rapid DNA loss?

Two possibilities can be envisioned: Intergenic sequences might be maintained by (1) selection on length rather than on the exact sequence (selective constraint hypothesis) or by (2) the balance between addition of DNA through large insertions and attrition of DNA through small deletions (dynamic equilibrium hypothesis). The selective constraint hypothesis predicts that intergenic regions should remain stable through time. The lengths of particular loci in extant species should be comparable not only to one another but also to that of the ancestral sequence. In contrast, the dynamic equilibrium hypothesis predicts that individual intergenic regions will go through large fluctuations in size, increasing sharply through large insertions and then continually shrinking from small deletions. Under this model, the maintenance of intergenic DNA would largely occur in aggregate across the whole genome and to a lesser extent at any region in particular.

With respect to the dynamic equilibrium hypothesis, selection on the function of genes will affect the rates of intergenic length evolution by eliminating any indel that removes a functional site (Ptak and Petrov 2002). As an intergenic sequence becomes shorter, for example, such constraint would retard further reduction in length. It is also possible that intergenic regions have both a minimum and a maximum length. When the length becomes very close to the minimum, increases in length from insertions may be promoted by positive selection. More complex phenomena may also be involved, with small deletions becoming slightly deleterious (and insertions slightly advantageous) as the length becomes too short. Conversely, if the intergenic region becomes too long, selection may promote fixations of deletions and retard fixation of insertions. The specifics of these processes are likely to vary significantly among different regions. The critical distinction from the selective constraint hypothesis is that the dynamic equilibrium model postulates that the lengths of intergenic regions may vary substantially between the possible low and high limits without strong impairment of function. The maintenance of the length between such boundaries may then be caused by the neutral or nearly neutral fixation of frequent but small deletions and rare but longer insertions.

The results presented in this paper are consistent with the dynamic equilibrium hypothesis. We documented the insertion of at least three approximately 600-bp transposable elements and one approximately 500-bp sequence of a numt in this region between 3 and 5 MYA in an ancestor of the *D. melanogaster* species complex. We saw no more insertions in this region since the diversification of the species complex, corresponding to a total of approximately 6.4 Myr of evolutionary time. The current lengths of this region in all of the species are shorter than they were in the ancestor, yet similar to each other because the attrition process has been occurring at similar rates for a similar amount of time. Also consistent with the dynamic equilibrium hypothesis, very little orthologous DNA has remained in all four species.

Although the pattern toward DNA loss is clear, we cannot distinguish the relative contributions of small deletions (<400 bp) versus larger ones to DNA attrition at this locus. Based on previous estimates of the rate DNA loss through small deletions in Drosophila, we expect that since 2.3 MYA, an unconstrained region should retain approximately 75% of its DNA. In comparison, the average amount of DNA retained at the studied locus in the four species since their MRCA is between 77% (based on the minimum estimate of the ancestral length of 1,847 bp) and 62% (based on the more likely estimate of 2,329 bp). Based on these estimates, there appears to be no reason to invoke the effect of deletions longer than 400 bp.

However, the small size of the studied intergenic region (~2 kb) and the requirement for a successful PCR and thus the presence of two priming sites approximately 1.2 kb apart has biased our observation against longer indels (Ptak and Petrov 2002). Any deletion larger than 2 kb would by necessity have been missed, although this bias is also present in all previous studies of indels in Drosophila (Pritchard and Schaeffer 1997; Ramos-Onsins and Aguade 1998; Robin et al. 2000; Blumenstiel, Hartl, and Lozovsky 2002; Petrov 2002*a*). Thus, it is entirely possible that deletions longer than 400 bp both occur with reasonable frequency and contribute to the length evolution of longer intergenic regions.

Our results demonstrate that at least some intergenic loci in Drosophila are substantially longer than the minimum allowable length and that their maintenance in this state may in part be mediated by the interplay between sporadic and long insertions and continuous but smaller deletions. The comprehensive study of the exact balance between mutational and selective forces in the maintenance of intergenic DNA will have to wait until the sequencing of multiple strains of *D. melanogaster* and its sibling species.

Acknowledgments

The authors would like to thank G. Gibson and S. Nuzhdin for their generous gifts of fly stocks. In addition, we gratefully acknowledge K. Ingram, D. Bensasson, and A. Hirsh for their valuable feedback on this manuscript and J. Davis and V. Kapitonov for their assistance regarding DNAREP1_DM analysis. The comments by the Associate Editor and the two reviewers substantially improved the manuscript.

Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt et al. (100 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

Bensasson, D., D.-X. Zhang, D. L. Hartl, and G. M. Hewitt. 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses. Trends Ecol. Evol. **16**:314–321.

Bergman, C. M., and M. Kreitman. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res. **11**:1335–1345.

Berry, A. J., J. W. Ajioka, and M. Kreitman. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129**:1111–1118.

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

Depaulis, F., and M. Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15**:1788–1790.

Hasegawa, M., and H. Kishino. 1990. Phylogenetic Inference from DNA Sequences. Fourth International Congress of Systematic and Evolutionary Biology. University of Maryland and The Smithsonian Institute.

Hilton, H., R. M. Kliman, and J. Hey. 1994. Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48**:1900–1913.

Jensen, M. A., B. Charlesworth, and M. Kreitman. 2002. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics **160**:493–507.

Kapitonov, V. V., and J. Jurka. 1999. DNAREP1_DM. Repbase Update Release 3. 4.

———. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. USA **100**:6569–6574.

Kumar, S., K. Tamura, and M. Nei. 1994. MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput. Appl. Biosci. **10**:189–191.

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

———. 2002*b*. Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. **61**:531–544.

Petrov, D. A., Y.-C. Chao, E. C. Stephenson, and D. L. Hartl. 1998. Pseudogene evolution in *Drosophila* suggests a high rate of DNA loss. Mol. Biol. Evol. **15**:1562–1567.

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

Powell, J. R. 1997. Progress and prospects in evolutionary biology: the *Drosophila* model. Oxford University Press, New York.

Pritchard, J. K., and S. W. Schaeffer. 1997. Polymorphism and divergence at a *Drosophila* pseudogene locus. Genetics **147**: 199–208.

Ptak, S. E., and D. A. Petrov. 2002. How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. Genetics **162**:1233–1244.

Ramos-Onsins, S., and M. Aguade. 1998. Molecular evolution of Cecropin multigene family in *Drosophila*: functional genes vs. pseudogenes. Genetics **150**:157–171.

Robin, G. C. D. Q., R. J. Russell, D. J. Cutler, and J. G. Oakeshott. 2000. The evolution of an alpha-esterase pseudogene inactivated in the *Drosophila melanogaster* lineage. Mol. Biol. Evol. **17**:563–575.

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

Sokal, R., and F. J. Rohlf. 1997. Biometry. W. H. Freeman (New York).

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–596.

Wang, W., K. Thornton, A. Berry, and M. Long. 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295**:134–137.

Adam Eyre-Walker, Associate Editor

Accepted November 18, 2003