Letter to the Editor

Pseudogene Evolution in *Drosophila* Suggests a High Rate of DNA Loss

Dmitri A. Petrov, Yu-Chan Chao,† Edwin C. Stephenson,‡ and Daniel L. Hartl*

Department of Organismic and Evolutionary Biology, Harvard University; and *Harvard University Society of Fellows; †Institute of Molecular Biology, Academia Sinica, Nankang Taipei, Taiwan, Republic of China; and ‡Department of Biological Sciences, University of Alabama

Pseudogenes—nonfunctional copies of functional genes—are very common in mammals, with many genes having tens or even hundreds of pseudogene copies (Weiner, Deininger, and Efstratiadis 1986), yet they are exceedingly rare in *Drosophila*, for which very few putative pseudogenes have ever been reported (Jeffs and Ashburner 1991). In addition, some *Drosophila* sequences originally described as pseudogenes (Jeffs and Ashburner 1991; Sullivan et al. 1994) were later demonstrated to be novel functional genes (Long and Langley 1993; Begun 1997). Nevertheless, despite the rarity of bona fide pseudogenes in *Drosophila*, its genome does harbor some nonfunctional sequences that appear to be unconstrained by selection and that evolve much like pseudogenes. One of the clearest examples of such DNA are the so-called ''dead-on-arrival'' (DOA) copies of non-LTR retrotransposable elements. These copies are generated frequently as by-products of transposition of active non-LTR elements. They lack 5′ sequences, including promoters and parts of open reading frames of proteins essential for transposition, and therefore they are usually predicted to evolve essential as pseudogenes. Recently, we were able to assess this prediction directly for at least one particular non-LTR element, *Helena*, in the *Drosophila melanogaster* and the *Drosophila virilis* species groups. Our approach relied on using maximum parsimony to separate the evolution of individual DOA insertions of *Helena* from the evolution of active lineages, which allowed us to demonstrate a lack of purifying selection acting on individual DOA elements (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1997, 1998).

The pattern of spontaneous substitutions observed in these unconstrained DOA sequences yielded a surprising result. We discovered a striking asymmetry in the pattern of spontaneous length substitutions. Not only were deletions found to outnumber insertions almost 9 to 1, but deletions were also much larger on average, ranging in size from 1 to 432 bp, with an average of 25 bp, while insertions ranged from 1 to 7 bp, with a mean of 2.8 bp. The preponderance of long deletions in DOA copies of *Helena* leads to very rapid loss of DNA from these sequences, more than 60-fold higher than that observed for mammalian pseudogenes. Although this high rate of DNA loss may, in principle, be a result of either biased mutation or selection for smaller genome size (Charlesworth 1996; Petrov, Lozovskaya, and Hartl 1996), we have been able to argue in favor of the mutational hypothesis by pointing out that the lengths of deletions are not positively correlated with the age of individual DOA *Helena* elements. Such positive correlation would be expected if the removal of DNA per se were selectively favored (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998).

If such a high rate of DNA loss is shared by most sequences in the *Drosophila* genome, it would help to explain the paucity of pseudogenes, which may be created just as frequently as in pseudogene-rich taxa but would be eliminated from the genome through rapid DNA loss much more quickly in *Drosophila*. It might also shed some light on the long-standing mystery of the C-value paradox (Thomas 1971) by suggesting that the vast differences in genome sizes among organisms may be due in part to the differences in the rate of loss of ''junk'' DNA.

On the other hand, the high rate of DNA loss may be a peculiar property of *Helena*. Nothing about the distribution of deletions and insertions in *Helena* suggests that the sequence of this element should be particularly prone to deletions. The possibility nevertheless remains that *Helena* may suffer a disproportionately high deletion rate as a result of either being recognized as a transposable element or being multiply repeated in the genome. It has now been firmly established that in many organisms, including *Drosophila* (Henikoff and Matzke 1997; Pal-Bhadra, Bhadra, and Birchler 1997; Selker 1997; Yoder, Walsh, and Bestor 1997), repeated sequences can be recognized and, in some cases, specifically inactivated, modified, and/or mutated. It has been hypothesized that the recognition and inactivation of repeated sequences may serve as a genomic defense mechanism against unchecked expansion of transposable elements (Bestor and Tycko 1996; Yoder, Walsh, and Bestor 1997). Given these precedents, it seems not out of the question that the high rate of deletions in *Helena* and other transposable elements may be due solely to their repetitive nature. If this is indeed the case, such a system of targeted deletion of repetitive DNA would represent a remarkable new strategy of genomic defense against invading DNA sequences—a defense that not only functionally inactivates these sequences, but also prevents their persistence and accumulation in the genome.

The possibility that bona fide *Drosophila* pseudogenes may experience a lower rate of DNA loss than

Abbreviation: DOA, dead on arrival.

Key words: spontaneous mutation, pseudogenes, *swallow* gene, DNA loss, *Helena*, C-value paradox.

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

Mol. Biol. Evol. 15(11):1562–1567. 1998

transposable elements such as *Helena* was highlighted recently by an investigation of molecular evolution of a pseudogene of *Larval cuticle protein* (*Lcpψ*) (Pritchard and Schaeffer 1997). Unlike *Helena*, *Lcpψ* appears to experience deletions and insertions at almost equal frequencies (six deletions and five insertions). Because deletions in *Lcpψ* are larger than insertions, the overall rates of DNA loss are similar in *Lcpψ* and in *Helena*. Nevertheless, the *Lcpψ* analysis is in line with the prediction of an altered ratio of deletions to insertions in transposable or multiply repeated sequences.

There is yet another possibility. Like many other transposable elements, most copies of *Helena* reside in pericentric heterochromatin (unpublished data). Since *Lcpψ* resides in euchromatin, the difference in the profiles of length mutations between *Lcpψ* and *Helena* may be a reflection of different mutational spectra in euchromatin versus heterochromatin. To obtain additional evidence bearing on whether the preponderance of large deletions is an exclusive property of multiply repeated, transposable, or heterochromatic DNA, we investigated molecular evolution of another euchromatic bona fide *Drosophila* pseudogene, *swallowψ* (*swwψ*) (Chao et al. 1991).

Chao et al. (1991) first described *swwψ* in the course of their analysis of the functional *sww* gene. The pseudogene is located immediately downstream of the functional copy of *sww* and appears to be a relatively recent direct duplication. The sequences of both *sww* and *swwψ* are deposited in GenBank under the accession number X56023.

Several features of *swwψ* suggest that it is not functional. First of all, it does not appear to be transcribed, since no cDNA clones corresponding to *swwψ* have been found, and RNAse protection assays fail to protect probes specific to *swwψ*. While *swwψ* does have a long open reading frame, it is missing any recognizable upstream regulatory sequences and the start codon. Furthermore, if *swwψ* were transcribed and translated, the *swwψ* protein would be missing 160 amino acids from its amino end, in addition to four gaps of 1, 3, and 16 amino acids in the body of the putative protein, and the protein would terminate prematurely compared with the protein sequence of the functional *sww* gene.

Comparison of the functional *sww* and *swwψ* nucleotide sequences (table 1 and fig. 1) showed a moderate proportion of nucleotide differences (8.5%) and a significant number (13) of insertions/deletions. The majority of indels result in the shortening of *swwψ* compared with *sww* (10 vs. 3); accordingly, *swwψ* is 15% shorter than *sww* (1,643 bp vs. 1,933 bp). The indels range in size from 1 to 138 bp, with an average size of 37 bp and a standard deviation of 45 bp.

When comparing two sequences, it is generally not possible to determine whether differences correspond to mutations in one sequence or the other. However, because we are comparing a functionally constrained sequence of *sww* with an unconstrained sequence of *swwψ*, we may be justified in assuming that most observed differences are due to mutations that have occurred in the pseudogene. We have the strongest grounds for making this assumption for mutations that would be likely to have a large detrimental effect if they occurred in *sww*, which is the case for indels and replacement substitutions in the coding region of *sww*.

Using this rationale, we infer that all eight indels in the alignment of coding regions of *sww* and *swwψ* correspond to deletions in *swwψ*. Thus, the observed ratio of deletions to insertions in *swwψ* is 8 to 0, which is entirely consistent with the pattern observed in DOA copies of *Helena* (87 deletions vs. 10 insertions) ($\chi^2 = 0.92$, $P = 0.34$). On the other hand, it is significantly different, although only marginally, from the pattern observed in *Lcpψ* (6 deletions vs. 5 insertions) by Pritchard and Schaeffer (1997) (χ^2 test, $P = 0.03$; χ^2 test with Yates correction for continuity, $P = 0.056$; Fisher's exact one-tailed test, $P = 0.04$).

The assumption that all differences between a functional gene and its pseudogene are due to substitutions in the pseudogene is valid only for strongly deleterious mutations. Thus, it is likely that some of the nucleotide differences between *sww* and *swwψ*, especially those in synonymous positions, correspond to substitutions in *sww*. The nonuniform distribution of point substitutions among the three codon positions is consistent with this prediction: among 100 nucleotide polymorphisms, 26 polymorphisms map to the first position, 29 map to the second, and 45 map to the third ($\chi^2 = 6.26$, $P = 0.04$). This asymmetry is probably due to stronger purifying selection at mostly nonsynonymous first and second codon positions compared with the mostly synonymous third positions, resulting in a larger proportion of substitutions in the first and second positions than in the third positions taking place in *swwψ*. In order to avoid gross overestimation of the total number of substitutions in *swwψ*, we estimated the number of substitutions in *swwψ* by first calculating the proportion of replacement substitutions (Jukes-Cantor one-parameter method, $K_n = 0.072$) and then by scaling K_n by the total number of positions in the alignment (1,149) to arrive at the estimate of 83 substitutions. The resulting proportion of nucleotide substitutions to the number of deletions (83 substitutions vs. 8 deletions) is consistent with the pattern observed for *Helena* (576 substitutions vs. 87 deletions) (*G*-test, $P = 0.22$). (Note that this is a conservative estimate of the deletion rate, since it is likely that some replacement substitutions occurred in *sww*).

The most striking feature of deletions in the DOA copies of *Helena* in *Drosophila* is that they are on average more than seven times larger than deletions in mammalian pseudogenes (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998). Because the rate of deletions in *Helena* is only 2.6 times as high as that in mammalian pseudogenes, it is primarily the vast difference in the average size of deletions that accounts for the 60-fold higher rate of DNA loss in *Helena* than in mammalian pseudogenes. It is therefore important to ascertain whether deletions in bona fide *Drosophila* pseudogenes are as large as those in *Helena*.

The eight deletions in the coding region of *swwψ* range in size from 3 to 138 bp, with an average length of 43 bp and a standard deviation of 51 bp. This is very

```
                                                               MET
sww      gttaattgtATTATTTCCCCCGCTTTTCGGATTTCCGCATAAAAAGCG    ATG AGT CTA CAG GAC GAG AGT TTT CCG
swwψ     ????????????????????????????????????????????????    ??? ??? ??? ??? ??? ??? ??? ??? ???

sww      ACG GAC GAG CTG TTT GAC CAG CTG AAC AAT TTG AGT AGC AGT GGC GCC AGG AAT ACC TGG TTC GCG
swwψ     ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ???

sww      GAG CAC CAT AAG CCC GCA GTC TTC GAG CGG GAT ACA GCG CCA TTT TTG GAG ATC TGC TAC GCG GAT
swwψ     ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ??? ?.. ... ... ... ... ... ... ... ...

sww      CCA GAC TTT GAT GCG GAT GGG GAT GTG GCC AAC AAG AGC GCC AAG ACA TGC GTA AGC GAT CCC GTG
swwψ     ..G ... ... ... ... --- --- ....... ... ... ... ... ... ... ... ... ... ... ... ... ...

sww      GGT CGT GAT CAG GAG GAT GAG GAC GAC TAT GAT GAG GAT GTC GAT G gtaagaagtg-----------ttcg
swwψ     ... ..C ... ... ... ... ... ... ... ... ... .G. ... ... ... . ........cttttccccctccccat

sww      ctctctctctacacttgcttaactacaatgggagaaatattcccatcaaacatatctcatata------------------------
swwψ     .c.a......ta.............a.c.a.........................t......cagggaactttttctgtgcacgtt

sww     -----------------------------------------------------------tgtataattttggtacttccatgtctcga
swwψ    accttttgttataacttcatacgctttcatttcaaaatttgtttcttaccattcataca.a..........a.........a.....c

sww     tgtataatgggtacttcattcaatgggcacagtcaattagctagtgctaacatgtatcctttcttattcaaccag  GC GAT GAT
swwψ    ....a.............ga.....t......c.....g..c.......................c.........  .T ... ...

sww     CAT AAA CTG GGT TGC GAG AAG GCT CCA TTG GGC AGC GGG CGC TCC AGC AAG GCG GTC TCT TAC CAG
swwψ    ... ... ..A ... G.. .T. G.. ... ... .C. ... ... ... ..G ..T ..A ... ... ... AAC ... .T.

sww     GAC ATC CAT TCG GCC TAC ACG AAG CGC CGC TTC CAG CAC GTG ACC AGC AAG GTG GGC CAG TAC ATA
swwψ    ... ... ..C ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

sww     GCG GAG ATC CAG GCG CAG GAC CAA AAG AGA CGC AAT GTG AAG TTC GCC GGA TTC CAG CGA GTG AAC
swwψ    ... ... ... A.. ... ... ... ... ... ..- --- --- --- --- --- --- --- --- --- --- --- --.

sww     TCT ATG CCG GAG AGT CTA ACG CCC ACA TTG CAG CAG GTG TAT GTC CAT GAT GGT GAC TTC AAG GTG
swwψ    GAC ... ... A.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

sww     GAC AAA AAC TGC CAG ACT CAC TCC AAC TCC GAT TCG AAT TAC AAT TCC AAT TCA AAC AAC TCT AGC
swwψ    ... ..G ... ... ... ... ... ... ... ... A.. ... ..C .G. ... ... ... ... ... ... ..C ...

sww     AGC AGC TTT GAT CGA TTG CTG GCC GAG AAC GAG AGC CTG CAG CAA AAG ATC AAC TCA TTG AGA GTA
swwψ    ... .A. ... ... ... ... ... ... ... ... ..T ..T ... ... ..G G.. ... G.G ... ... .A. .C.

sww     GAA GCG AAG CGT CTG CAG GGC TTC AAC GAG TAT GTC CAG GAA CGA CTG GAC AGA AAG ACA GAT GAT
swwψ    ..G ... ... ... A.. G.. ..G ... ... .T. ... C.. ... ... ... ... A.. ... ... .G. ... ...

sww     TTT GTG AAG ATG AAG TGC AAT TTC GAG ACC CTG CGC ACC GAG CTA AGC GAA TGC CAG CAG AAG CTT
swwψ    .A. ... .C. ..T ... ... ... ..T ... ... ..A ... ... ... ... ... AT. ... ... T.T .G. ..G

sww     AGG CGC CAG CAG GAC AAC TCA CAG CAC CAC TTC ATG TAC CAC ATT CGA TCG GCG ACC AGT GCC AAG
swwψ    ..T ... ... A.. ... ... ... ... .C. ... ... ... ... ... ... ... ... .-- --- --- -AA ...

sww     GCC ACT CAA ACG GAT TTC CTG GTG GAC ACC ATA CCC GCC TCC GGA AAC GTC CTG GTC ACA CCC CAT
swwψ    ... ... ... ... ... A.. ... ... ... G.. ... ... ..G ... .A. ... ... ... ... ... ... ...

sww     CCC CTG GGC GAC CTG ACC TAC AAC AGC AGC AAA GGA TCC ATC GAG TTG GCA CTG CTC AGT GTG GCG
swwψ    ... ..T ... ... ... ..- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --.

sww     CCT TCT GCC CGA GTG GCC CAG AAT CCC GTC CAG GTC CAA CGC GCG ATT CAT CCA CAA TCT TTG GAC
swwψ    ... A.. ... ... ... ... ... ... ... ... ... ... ... ... ... ..T ... ..C ..C ... ..C ..A ...

sww     TTT AGC AGC GTT AGC ACC GAA GCT GAT GGC AGC GGT AGT Ggtaagcaatc-gcaatgctattttattagatctg
swwψ    ... ... ... ... .AA ..G ... ... A.. .C. ... ... ... ...........ctat........ga..a.......
```

FIG. 1.—The alignment of *sww* and *swwψ*. Exons are shown in uppercase letters, and introns are shown in lowercase letters. The coding region is shown in three-nucleotide blocks, with the translation start site identified with MET and the stop codon with STOP. Dots identify positions in which the sequence of *swwψ* is identical to that of *sww*, and dashes show the inferred positions of deletions. Question marks identify the sequences in *sww* that are absent from *swwψ*. *swwψ* is missing all recognizable upstream regulatory signals, the 5′ UTR, and the first 43 codons. The beginning of the *swwψ* sequences is 140 nucleotides downstream of the second of two polyadenylation sites in *sww*. The downstream limits of *swwψ* have not been determined.

```
sww       tagattatttcgaataagattgactaacatggatttttgttggtttctattcaag  GC GAA CAT CGT GTG GAA ACC TCA
swwψ      ..........-------------------------------------........  .. ... ... ... ... ..C ... ...

sww       GCC TCC AGG TTG GTC AGA AGA ACC CCG GCG CCC AAC AAC TCG GAA ACC AGC CAG CCG AGC AGC AAC
swwψ      ..G ... ... ... .TG ... ... ... ... --- ... .T. --- --- --- --- --- --- --- --- --- ---

sww       GAC TCG GCC ATC GAG GTG GAG GCG CAC GAG GAG GAG CGA CCC AGC TCC AGG CGG CAG TGG GAA CAA
swwψ      --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

sww       CAG GGG GAG CTC ATC TCG CCC AGG CAA TGG GGC CAG CAT GAG GGC ATG TAC TAC TTT GAC AAG CGC
swwψ      --- --- -.T TCA ... ..C T.. ..C ... ... ... ... ... ..C ... ... ... ... ... ... ... ...

sww       AAC AAC CGA GTC ATC GAG GTG ATG GGC TTC AAT ATC AGT CAG GGG CGC AAT CAG AGC CAT GAC ACC
swwψ      ... ... ... ... ... ..A .-- --- --. ... ... ... ... ... C.. A.. ... ... ... ..A ... ...

sww       ATT CAT AAT CAG AGC ATC AAC GAT AGT CAG ACG CGT CTG CTG GTC CAC TCG ATG TCG ATG TCG CAT
swwψ      ... ... ... ... ... ... .G. ... ... ... .A. ... T.. ... ... .G. ... ... ... ... --- ---

sww       TTG GAG GCG CAT GAC CAC TTT AGG AGT AAA AGG ACG ACA CTG GGC AGT CGG ATG CTA CGA TTC CTG
swwψ      --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

sww       GGG CCC TGC GTT CGC TGC CGT AAT GGT GAT CCA TTG AAC CGC AGC AAT GTC ACA TAC AAG GAT GGT
swwψ      --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
                                                                          STOP
sww       TTG CCT GCG ATG CCC GAG GAG GAG TTC GTT GAC CAA AGG AAC CAG CGC TAG TCCAGCCTTCCACCATCCA
swwψ      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...................

sww       CTTTTATTTATTTATTGTATGTATTATCATATCCTGTTATTTTATGTCGCGTTATGTCGATTCAATTAGTCCAAATTTTTAATGCCA
swwψ      ........................A...........C........CA.......A..-...C.........G...............

sww       TAGCATTCAGTTAACCAATATGTGCATGTATAACCAAATTT
swwψ      C........................................
```

FIG. 1 *(Continued)*

similar to the pattern observed in the *D. melanogaster* subgroup *Helena* data set (Mann-Whitney two-tailed *U* test, $P > 0.05$). In the case of DOA elements from the *D. melanogaster* subgroup, deletions range in size from 1 to 432 bp, with a mean of 34 bp and a standard deviation of 65 bp. Admittedly, because the overall number of deletions in *swwψ* is small, the power of comparison of the size distributions is low. But we can get a sense that these distributions are similar. In both cases, about half of all deletions are smaller than 10 bp—34 of 64 deletions in the *D. melanogaster Helena* data set (57%) and 4 of 8 deletions in *swwψ* (50%). Similarly, both distributions have a long right-hand tail. Importantly, the deletions in *swwψ* are at least as large as they are in *Helena* and occur at a similar rate when measured relative to the rate of point substitutions, indicating a similar rate of DNA loss. Indeed, based on the estimates of DNA loss from the *D. melanogaster Helena* data (Petrov and Hartl 1998), we would predict that the coding sequence of *swwψ* should be reduced by 22%, and, in fact, it is reduced by 23%.

Approximately 50% of deletions in both *Helena* data sets are flanked by short direct duplications of 2–7 bp in length, suggesting a homology-dependent mechanism of deletion formation, such as recombination or DNA replication slippage (Petrov, Lozovskaya, and Hartl 1996; Petrov and Hartl 1998). The same is true for deletions in *swwψ*. Two of eight deletions are flanked by direct repeats of 2–4 bp (data not shown). Also, as is the case for deletions in *Helena*, there is no evidence of correlation between the presence or absence of direct duplications at the termini of a deletion and the deletion size.

Thus, it seems that the patterns of deletions and insertions in *swwψ* and in DOA copies of *Helena* are completely consistent with each other. Deletions in *swwψ* are frequent and large and significantly outnumber insertions, suggesting that the preponderance of large deletions in DOA copies of *Helena* is not an exclusive feature of multiply repeated, transposable, or heterochromatic DNA.

In addition to our analysis of *swwψ*, a recent molecular analysis of *Adh* retrosequences in the *D. obscura* species group (Luque, Marfany, and Gonzales-Duarte 1997) revealed a pattern of deletions and insertions very similar to those of *Helena* and *swwψ*. The authors observed 14 deletions, ranging in size from 1 to 34 bp, with an average of 10.1 ± 12.2 bp, and 1 insertion of 6 bp. Similar to the distribution of deletion sizes in *Helena* and *swwψ*, approximately half are smaller than 10 bp (seven deletions of 1 bp, one of 2 bp, and one of 6 bp), and the rest are significantly longer than 10 bp (two of 22 bp and one each of 20, 28, and 34 bp). Because *Adh* retrosequences in the *D. obscura* group may not be evolving as pseudogenes, the observed pattern of deletions and insertions may reveal not only the spontaneous profile of mutations in these sequences, but also the action of natural selection. For instance, all but one deletion occur outside the open reading frame, suggesting that purifying selection has been acting to preserve the coding capacity of the *Adh* retrosequences. However, unless natural selection in *Adh* retrosequences strongly

Table 1
Sequence Comparison of *sww* and *swwψ*

			Indels	
Region	Nuc[a]	Jukes-Cantor Distance	Add DNA to *sww*	Add DNA to *swwψ*
Exons			8	0
Synonymous	242.5	0.147 ± 0.027	—	—
Nonsynonymous	906.5	0.072 ± 0.009	—	—
Introns	227	0.124 ± 0.025	1	3
3′ untranslated region (3′ UTR)	146	0.057 ± 0.020	1	0
Introns and 3′ UTR	373	0.097 ± 0.017	2	3

[a] Number of positions in the comparison.

favors deletions over insertions, it is more likely that it is the bias in the mutational spectrum that is responsible for the preponderance of deletions in the retrosequences of *Adh* in the *D. obscura* group.

Based on the *swwψ* and *Adh* data, we would conclude that the preponderance of relatively large deletions among length mutations, as well as the high rate of DNA loss exhibited by DOA copies of *Helena*, is likely to be a general property of mutation in *Drosophila*. But what about *Lcpψ*? Pritchard and Schaeffer (1997) reported a ratio of six deletions to five insertions in *Lcpψ*, which is significantly different from the pattern in *Helena* (87 deletions and 10 insertions), that of retrosequences in *Adh* (14 deletions and 1 insertion), and that in *swwψ* (8 deletions and 0 insertions).

We have no ready explanation of why *Lcpψ* appears to be different. We would emphasize, however, that the only difference between the length mutations in *Lcpψ* and those in *Helena* or *swwψ* lies in the higher frequency of insertions observed in *Lcpψ*. The pattern of deletions in *Lcpψ* is not markedly different from that in either *Helena* or *swwψ*. In particular, the ratio of the number of fixed point substitutions (estimated using the Jukes-Cantor one-parameter method) and the number of deletions in the comparison of the *D. simulans* and *D. melanogaster Lcpψ* genes is 58 to 4. This ratio is not statistically different from that of *swwψ* (*G*-test, $P = 0.58$) or that of *Helena* (*G*-test, $P = 0.10$). Furthermore, deletion sizes in *Lcpψ* also appear to follow the distribution of deletion sizes in *Helena* and *swwψ*. At 9.2 ± 11.6 bp, the average size of deletions in *Lcpψ* is somewhat smaller than that of *Helena* or *swwψ* deletions, but this difference is not statistically significant (for both comparisons, Mann-Whitney two-tailed *U*-test, $P \geq 0.05$). Moreover, the general pattern of approximately half of all deletions being smaller than 10 bp (four of six) with some deletions being much larger (in this case, 22 and 26 bp) is preserved in *Lcpψ*. The size distribution of insertions in *Lcpψ* is also very similar to that observed for *Helena*, for retrosequences of *Adh*, and for *swwψ*, with five deletions ranging in size from 1 to 3 bp. Therefore, despite the higher frequency of insertions in *Lcpψ*, the much larger average size of deletions relative to that of insertions still ensures a high rate of DNA loss.

All of the unconstrained sequences studied to date in *Drosophila*, both heterochromatic and euchromatic, transposable and nontransposable, multiply duplicated and not, seem to experience similar high rates of DNA loss. This pattern is inconsistent with the hypothesis of a high rate of DNA loss being a genomic defense strategy employed by organisms exclusively against multiply repeated invading sequences. It is consistent, however, with the idea that this high rate of DNA loss is a general property of the mutation in *Drosophila*. Nevertheless, much more information on the mutational patterns of different DNA sequences in different regions of the genome must be gathered before reliable estimates can be obtained of both the average profiles of length mutations and the extent to which they may vary from one region or type of DNA sequence to the next.

Acknowledgments

We thank M. Siegal, R. Lewontin, J. Morris, C.-T. Wu, and members of the Daniel Hartl laboratory for helpful discussions. We are also thankful to J. K. Pritchard and S. W. Schaeffer for sharing their results prior to publication and for helpful comments. This work was supported by the Junior Fellowship from the Harvard University Society of Fellows to D.A.P., NIH grants GM33741 and HG01250 to D.L.H., and NSF grant IBN-9630705 to E.C.S.

LITERATURE CITED

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

Bestor, T. H., and B. Tycko. 1996. Creation of methylation patterns. Nat. Genet. **12**:363–367.

Chao, Y.-C., K. M. Donahue, N. J. Pokrywka, and E. C. Stephenson. 1991. Sequence of *swallow*, a gene required for the localization of *bicoid* message in *Drosophila* eggs. Dev. Genet. **12**:333–341.

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

Henikoff, S., and M. A. Matzke. 1997. Exploring and explaining epigenetic effects. Trends Genet. **13**:293–295.

Jeffs, P., and M. Ashburner. 1991. Processed pseudogenes in *Drosophila*. Proc. R. Soc. Lond. B Biol. Sci. **244**:151–159.

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.

Luque, T., G. Marfany, and R. Gonzales-Duarte. 1997. Characterization and molecular analysis of *Adh* retrosequences in species of the *Drosophila obscura* group. Mol. Biol. Evol. **14**:1316–1325.

Pal-Bhadra, M., U. Bhadra, and J. A. Birchler. 1997. Cosupression in *Drosophila*: gene silencing of Alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent. Cell **90**:479–490.

Petrov, D. A., and D. L. Hartl. 1997. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila.* Gene **205**:279–289.

———. 1998. High rate of DNA loss in the *D. melanogaster* and *D. virilis* species groups. Mol. Biol. Evol. **15**:293–302.

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

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

Selker, E. U. 1997. Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion. Trends Genetics **13**:296–301.

Snyder, M., M. Hunkapiller, D. Yuen, D. Silbert, and J. Fristrom. 1982. Cuticle protein genes of *Drosophila:* structure, organization and evolution of four clustered genes. Cell **29**:1027–1040.

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.

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.

Yoder, J. A., C. P. Walsh, and T. H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. **13**:335–340.

Pierre Capy, reviewing editor

Accepted August 6, 1998