Genetica **115:** 81–91, 2002.

DNA loss and evolution of genome size in *Drosophila*

Dmitri A. Petrov
Department of Biological Sciences, Stanford University, 371 Serra St., Stanford, CA 94305, USA (Phone: (650) 736 1169; E-mail: dpetrov@stanford.edu)

Key words: deletions, insertions, non-LTR elements, numts, spontaneous mutation

Abstract

Mutation is often said to be random. Although it must be true that mutation is ignorant about the adaptive needs of the organism and thus is random relative to them as a rule, mutation is not truly random in other respects. Nucleotide substitutions, deletions, insertions, inversions, duplications and other types of mutation occur at different rates and are effected by different mechanisms. Moreover the rates of different mutations vary from organism to organism. Differences in mutational biases, along with natural selection, could impact gene and genome evolution in important ways. For instance, several recent studies have suggested that differences in insertion/deletion biases lead to profound differences in the rate of DNA loss in animals and that this difference *per se* can lead to significant changes in genome size. In particular, *Drosophila melanogaster* appears to have a very high rate of deletions and the correspondingly high rate of DNA loss and a very compact genome. To assess the validity of these studies we must first assess the validity of the measurements of indel biases themselves. Here I demonstrate the robustness of indel bias measurements in *Drosophila*, by comparing indel patterns in different types of nonfunctional sequences. The indel pattern and the high rate of DNA loss appears to be shared by all known nonfunctional sequences, both euchromatic and heterochromatic, transposable and non-transposable, repetitive and unique. Unfortunately all available nonfunctional sequences are untranscribed and thus effects of transcription on indel bias cannot be assessed. I also discuss in detail why it is unlikely that natural selection for or against DNA loss significantly affects current estimates of indel biases.

Introduction

Mutation is the central player in the Darwinian theory of evolution – it is the ultimate source of heritable variation, providing the necessary raw material for natural selection. Mutation is assumed to create heritable variation that is abundant, random, and undirected. Natural selection then directs evolution toward adaptive ends by sorting the initially random variants according to their adaptive values. Such a view is justified if we are interested in adaptation of organisms to their environment, because mutation cannot be biased toward phenotypes matching the ever-changing environment. However, there is abundant evidence that mutation is not truly random – some mutations are more common than others, independently of natural selection. The effect of this non-randomness on the ultimate course of evolution can be very important, yet it is very poorly understood [1, 2].

One key question that needs to be addressed is the extent to which mutational patterns vary in different taxa. Significant and persistent differences in mutational patterns could drive genetic and phenotypic divergence among species. Of course, even if mutational biases do vary a great deal across taxa, these biases could have only a minor effect on evolution. If natural selection is extremely strong and there exists only a single, stable, sharp major fitness peak, then even less abundant, but selectively favored variants should eventually come out on top in the evolutionary game. Thus, the second key question is whether the evolution of genes and genomes is, in fact, responsive to changes in mutational biases. The answer to this second question would give us a much better understanding of the importance of mutational biases, and, in general, of the role that biases in the introduction of genetic variation play in evolution [1].

Our current knowledge of mutational parameters is extremely limited. This is primarily because it is very difficult to study spontaneous mutation, especially in non-model organisms. The problem is twofold: on one hand spontaneous mutation is generally too infrequent to be investigated directly in the laboratory, whereas the inference of mutational patterns from variation in populations is hampered by the confounding influence of natural selection. Traditionally this problem has been addressed through the use of pseudogenes [3, 4]. Pseudogenes – nonfunctional copies of functional genes – are presumed to evolve without functional constraints, and therefore patterns of substitution in pseudogenes are assumed to reflect patterns of spontaneous mutation faithfully. While the pseudogene approach is very powerful, its utility is severely limited because in most organisms pseudogenes are not available. Some well-studied organisms, such as *Drosophila*, have very few pseudogenes, whereas other, non-model organisms have not been studied in sufficient molecular detail to provide a large enough number of pseudogenes for the analysis.

Although genomes of many organisms lack large numbers of pseudogenes, most of them do contain other kinds of unconstrained sequences. Practically all eukaryotic genomes contain defunct copies of transposable elements (TEs) and nonfunctional insertions of organellar DNA [5]. In addition, unlike bona fide pseudogenes, these sequences can be easily identifiable even in poorly studied organisms. Nonfunctional copies of TEs and insertions of organellar DNA can, therefore, provide an important new source of unconstrained nuclear DNA that can be used to study mutational patterns in diverse organisms.

In our work we have concentrated mostly on a particular kind of transposable elements, non-LTR (long terminal repeat) retrotransposons, most copies of which are 5′ truncated, non-functional elements that are predicted to evolve without functional constraint immediately upon their transposition (they are DOA – 'dead-on-arrival'). We have used a *Drosophila* non-LTR element *Helena* to show that *Drosophila* has a much higher rate of substantially longer deletions than mammals and some other insects [6–9]. Remarkably, it appears that the rate of DNA loss through the imbalance of small deletions and insertions is a good predictor of genome size and may be one of the key parameters in genome size evolution [10].

These intriguing conclusions are predicated partly on the validity of inferring mutational spectra in *Drosophila* genome from the study of DOA non-LTR elements. However mutation in non-LTR elements may have a specific, non-representative pattern, different from that in the rest of the genome. Non-LTR elements have a particular DNA sequence, they are transposable, repetitive, and may be located disproportionately in particular genomic regions. Do any of these properties result in significantly different mutational patterns in non-LTR element derived DNA? In this paper I will attempt to answer this question by reviewing indel spectra found in several *Drosophila* pseudogenes and in *D. melanogaster* nuclear insertions of mitochondrial DNA (numt). The totality of the data strongly suggest that much of the nonfunctional DNA in *Drosophila* displays a very similar, essentially indistinguishable pattern of indels resulting in the genome-wide rampant DNA loss.

Results and discussion

Using non-LTR retroelements to study spontaneous mutation

The rationale for using non-LTR elements as pseudogene surrogates has been explained in detail elsewhere [6, 11]. This approach takes advantage of the life-cycle of non-LTR elements, which has been elucidated in a variety of organisms [12–18].

The basic transpositional cycle of non-LTR elements is shown in Figure 1. The active element is first transcribed from an internal promoter, followed by reverse transcription and insertion of the resulting cDNA at an ectopic location [13]. Figure 1 shows a common result of transposition, whereby reverse transcriptase (RT) falls off the template prematurely, resulting in the insertion of a 5′ truncated copy. These 5′ truncated copies are inactive and DOA – they are essentially pseudogenes.

If we could follow all of the molecular changes in a DOA element back in time, the most recent changes would be in the DOA copy itself, suffered after it was created through abortive transposition. These changes would reflect evolution unconstrained by natural selection for the ability to transpose. Further back in time, however, molecular changes in the DOA element would coalesce with the active lineage that gave rise to it and those changes would reflect natural selection for transpositional competence.

To estimate patterns of mutation unaffected by selection for transpositional competence, first we need to separate the unconstrained evolution of the DOA copies from the constrained evolution of the active

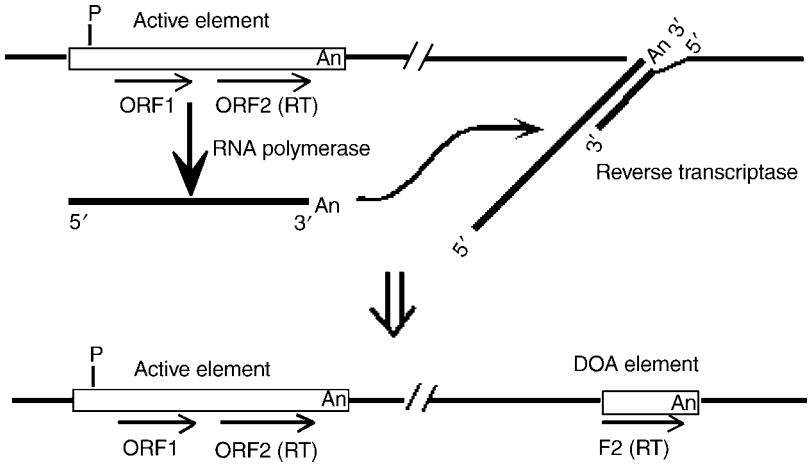

Figure 1. Transposition of DOA non-LTR elements. DOA elements are commonly created when the reverse transcriptase (RT) prematurely terminates reverse transcription of the mRNA produced by the active non-LTR element from its internal promoter (P).

lineages. The basic insight is that this task can be accomplished through the phylogenetic analysis of multiple copies of a non-LTR element. If the sampling is dense enough, such that the sample contains multiple elements per any active lineage, then the terminal branches on a phylogenetic tree would predominantly represent pseudogene-like evolution of DOA elements, while the internal branches would track evolution of active lineages. This is because changes that occur in active lineages have a chance of being incorporated in multiple DOA elements and thus would be classified as shared changes mapping to the internal branches. Mutations in the DOA lineages, on the other hand, are independent, and barring parallel mutation, would be unique and map to the terminal branches.

Does the non-LTR element based approach work in practice?

To test this approach we gathered a dataset of multiple copies of non-LTR element *Helena* [19] from the *D. virilis* and *D. melanogaster* species groups [6, 7]. We tested the predictions of our approach by classifying all substitutions in the RT coding region as affecting either the 1st, 2nd, or 3rd position of the codons. Since many substitutions in the 3rd position are synonymous, and most substitutions in the 1st and 2nd position are nonsynonymous, we expected the terminal-branch substitutions to be random in all three positions, whereas the majority of substitutions in constrained sequences along the internal branches should be in the 3rd position.

This was exactly what we found in the *D. melanogaster* [7]: the terminal branches show equal rates of 1st, 2nd, and 3rd codon position changes ($P = 0.61$), whereas the internal branches show a highly significant excess of third position changes ($P = 2 \times 10^{-7}$). We observed a similar pattern in the *D. virilis* dataset [6]. In addition, *Drosophila Helena* RT sequences contain a number of deletions and insertions (indels), which are likely to be deleterious and therefore should map onto unconstrained, terminal branches. Most of the indels do indeed map to the terminal branches (83 out of 87 deletions and 8 out of 9 insertions).

If the terminal branches do correspond to the pseudogene evolution of the DOA copies after their creation, then the length of each branch should be proportional to the age of the individual DOA elements. Since all types of substitutions should accumulate with time, we expected to observe a positive and monotonic correlation between the numbers of indels and point substitutions along each terminal branch.

This is indeed what we observed (Figure 2): there is a strong positive correlation between the numbers of deletions and point substitutions in the *D. virilis* dataset (sign test, $P = 0.043$) and in the *D. melanogaster* dataset ($P = 0.008$, Friedman's method for randomized blocks). (There is no detectable correlation between numbers of insertions and point substitutions, probably due to the small number of insertions – 1 and 8 insertions in the *D. virilis* and the *D. melanogaster* data, respectively.)

In total, the observations strongly argue that the terminal branches in the *Drosophila* data correspond to the pseudogene phase in the life-cycle of non-LTR

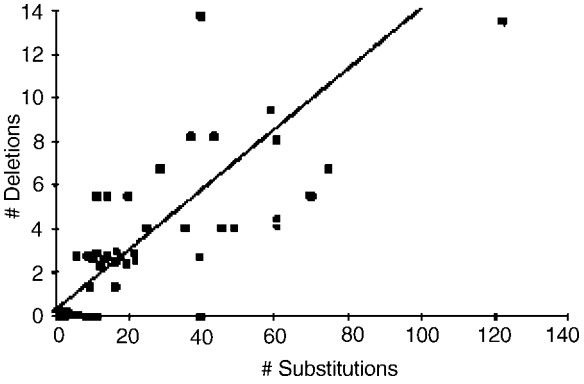

Figure 2. The combined data on the correlation between the rate of nucleotide substitutions and deletions in the DOA copies of *Helena* in *D. virilis* and *D. melanogaster*.

elements. Thus we are justified in using the substitutions along the terminal branches as estimates of spontaneous mutation.

High rate of DNA loss in Drosophila

The most striking feature of the mutation pattern in DOA copies of *Helena* is the extremely high rate of spontaneous DNA loss due to frequent and long deletions: deletions in *Helena* are on average 7 times longer and 3 times more frequent than they are in mammalian pseudogenes [6, 7, 20–23]. Insertions, on the other hand, are very short and similar in size and frequency in *Helena* and mammalian pseudogenes. This difference in indel spectra leads to a 20-fold higher rate of DNA loss per point substitution in DOA copies of *Helena*, which, combined with a 3-fold higher rate of point substitution per year [24], results in an approximately 60-fold faster loss of DNA in absolute time.

If these results hold for the *Drosophila* genome in general, such rampant DNA loss would result in very short persistence of nonfunctional DNA in the *Drosophila* genome. It can be calculated that a pseudogene fixed in a *Drosophila* lineage should lose half of its DNA in ∼14 Myr, compared to ∼880 Myr expected on average in mammals. This could be part of the reason why *Drosophila* has so few pseudogenes – even if they are formed at the same rate as in mammals, their DNA is lost very quickly leading to a low steady-state number of recognizable pseudogenes at any given time.

The *Drosophila* data suggest that the rate of spontaneous DNA loss might be an important parameter in the long-term evolution of genome size. Other things being equal, a high rate of DNA loss should result in compact, 'junk'-free genomes, whereas a slow loss of DNA would lead to large, 'junky' genomes. Of course, other things are not always equal, with other forces possibly affecting genome size to a greater degree. Nevertheless, we can establish the relative importance of spontaneous DNA loss by measuring the rate of DNA loss and genome size in a number of distantly related but similarly complex organisms and determining whether the rate of DNA loss negatively correlates with genome size. In such an analysis, other determinants of genome size are expected to appear as noise, adding scatter to the correlation. The presence of a detectable correlation would serve as clear evidence for the importance of deletion/insertion spectra in genome size evolution.

As the first such test of the DNA loss hypothesis, we have assayed the rate of DNA loss in Hawaiian crickets (Genus *Laupala*) and grasshoppers (Genus *Podisma*) [8, 9]. (In Laupala we assayed DNA loss using a new non-LTR elements (*Lau1*) cloned for our study [8], and in *Podisma* grasshoppers we used nuclear insertions of mitochondrial DNA (numts) [9, 25]. Genome sizes in these organisms span two orders of magnitude with *Drosophila* (165 Mbp) having the smallest genome size, *Laupala* crickets having an 11-fold larger genome (1910 Mbp), and *Podisma* grasshoppers having an almost 10-fold larger genome yet (18,150 Mbp). The *a priori* prediction tested in these studies was that the rate of DNA loss should be negatively correlated with the genome size.

Table 1 shows that the experimental data supported these predictions. There is strong negative correlation between rate of DNA loss and genome size. This negative relationship extends all the way from the compact *Drosophila* genome to the extremely large genomes of grasshoppers. Although the amount of data are quite limited, the correlation between the rate of DNA loss through small deletions and genome size is statistically significant ($P = 0.01$ for the log/log transformed data). It is important to note that the predictions of the slower rate of DNA loss were made *a priori*, exclusively on the basis of genome sizes, and were then confirmed experimentally by measuring the rates of DNA loss.

The key assumption here is that the rate of DNA loss measured in DOA elements [7, 8] or numts [9] reflects the mutational imbalance of deletions and insertions across all (or most) of the sequences in the assayed genomes. This may or may not be true. To fully understand the role that mutational DNA loss

Table 1. Rate DNA loss and genome size in insects

	Drosophila[a]	*Laupala* crickets[a]	*Podisma* grasshoppers[b]
Genome size (Mbp)	179	1910	18150
Rate of DNA loss (bp/per 1 bp substitution)	3.2	0.34	0.06
Half-life of a pseudogene (Myr)	14	615	880[c]–3500[d]

[a]Data from [8].
[b]Data from [9].
[c]Assuming the same absolute rate of nucleotide substitutions as in *Drosophila*.
[d]Assuming the same absolute rate of nucleotide substitutions as in *Laupala*.

may play in the evolution of genome size, it is essential to validate this assumption experimentally. In the remainder of this paper, I will discuss multiple possible sources of bias in our estimates of DNA loss and attempt to demonstrate that at least in *Drosophila* and at least in the case of the DOA non-LTR elements we can trust the estimates of indel biases. The experimental validation of the use of numts and of DOA non-LTR elements in other organisms still remains to be carried out.

When do terminal branches fail to trace the unconstrained evolution of DOA non-LTR elements?

Although the terminal branches in the *Drosophila Helena* datasets [6, 7] seem free of purifying selection, this does not mean that all substitutions along the terminal branches correspond to pseudogene evolution. In fact, part of the terminal branch leading to the most recently transposed element will always correspond to that part of the active element evolution, that occurs after transposition of the second most recent DOA element. This should not be a problem, as long as enough pseudogene elements per active lineage are sampled, so that the presence of a small amount of active element evolution along the most recent branch is swamped out by the large number of pseudogene substitutions along multiple and longer terminal branches leading to the older elements. We do not know *a priori* how many active lineages coexist in a species and thus cannot guess how dense the sampling needs to be. Therefore, we employ an empirical way of assessing whether the sampling is dense enough: enough elements need to be sampled so that the pattern of point substitutions along terminal branches shows no signs of purifying selection.

To illustrate this point we culled the *D. virilis* data into random sets of four, six, or 12 sequences and compared it to the full set of 18 sequences [11]. With only four sequences, purifying selection was evident in the strong preponderance of 3rd codon position substitutions in the terminal branches ($P = 5.5 \times 10^{-5}$), however, with 12 sequences this imbalance was no longer significant ($P = 0.051$), and was entirely absent with 18 sequences ($P = 0.18$). This exercise shows that in the case of *Helena* in the *D. virilis* species group having 18 sequences was sufficient, however, it does not ensure that it would not be necessary to collect more sequences for different elements in other species.

When do internal branches fail to trace the evolution of active lineages?

In our discussion, we tacitly assumed that all elements in the sample transposed independently (all copies are paralogous), which is why we could argue that shared substitutions among any two DOA elements must correspond to the evolution of an active lineage that produced them. However, it is clearly possible to resample the same element, present in the same allelic position, multiple times. The resampling of orthologous elements in different species will turn some of the internal branches into pseudogene branches. As long as we are interested only in pseudogene evolution and limit ourselves to the substitutions along the terminal branches, such a scenario would have no effect on our data. In addition, if the pseudogene internal branches can be identified, they can be used to collect data on mutational patterns as well [7].

Could DOA non-LTR elements be mobilized in trans*?*

Another tacit assumption behind our analysis is that DOA non-LTR elements are marooned and cannot transpose. Even though DOA elements cannot mobilize in *cis*, there is a possibility of another active element in the genome mobilizing DOA elements in *trans*. In fact, *trans* complementation of non-LTR

elements has been shown to occur [26, 27]. However, these studies also suggest that trans-mobilization should be very rare and limited to the very actively transcribed sequences. Because the vast majority of DOA elements are 5′ truncated and have no promoter, high levels of transcription are very unlikely. In addition, the clear evidence of purifying selection acting along internal branches in the *Drosophila Helena* data [6, 7] and in the dataset of the rodent *L1* elements [12], also implies that only active elements can efficiently transpose.

What if mutation in DOA elements is mostly caused by low fidelity transcription and reverse transcription?

Transposition of non-LTR elements proceeds through two mutagenic stages: transcription and reverse transcription. How can we be sure that mutations along the terminal branches are not substantially caused by these processes? To test for this possibility we sequenced six newly transposed copies of *Helena* [28], mobilized 10 years ago in a *D. virilis* hybrid dysgenic cross [19, 29]. We found no nucleotide substitutions or indels in these elements, giving us an upper bound of 7×10^{-4} substitutions per nucleotide generated as a result of transposition. This result shows that mutations occurring at transposition exert at most a minor effect, given that the average proportion of substitutions in the *Helena* data is more than 25-fold higher ($\sim$0.02 substitutions per nucleotide) than the upper bound of transposition-induced mutation. This is not that surprising given that each DOA copy undergoes only a single round of low fidelity transcription and reverse transcription compared to hundreds of thousands or even millions of rounds of DNA replication. DNA replication simply has many more chances to produce mutations compared to transcription and reverse transcription.

Indel patterns in other types of unconstrained DNA

Even if we correctly infer the pattern of mutations in DOA copies of non-LTR elements, it is difficult to extrapolate from non-LTR elements to other sequences in the genome. Non-LTR elements are peculiar in a number of ways. They are both repetitive and transposable. They may be preferentially located in particular parts of the genome (e.g., heterochromatin v.s. euchromatin). They may simply happen to have a peculiar sequence-specific mutational bias. All of these features may have an effect on indel patterns. The only way to assess this possibility is to study other unconstrained sequences in the genome that do not share all or some of the specific features of non-LTR element derived DNA.

Bona fide pseudogenes

Although bona fide pseudogenes are much rarer in *Drosophila* than in mammals, there are now four examples of well-established bona fide pseudogenes, whose evolution has been studied in detail [30–33]. These pseudogenes are distinct from DOA copies of *Helena* in a number of ways. They are non-homologous to *Helena*, they are not transposable, and all of them are located in euchromatin (one on the X chromosome and three on the third chromosome). Study of the indel bias in these pseudogenes can tell us whether the high rate of DNA loss in DOA copies of *Helena* is due to any one of those distinguishing features.

Table 2 summarizes the findings. In all cases the indel pattern is strongly biased toward DNA loss. In three cases out of four (except *Larval cuticle protein* ψ) the indel pattern is indistinguishable from that observed in *Helena*. (*Lcp*ψ has a higher rate of insertions, but not different size distributions, than do *Helena* and the other three pseudogenes.) In all four cases the rate of deletions per point substitution is not significantly different from that observed in DOA copies of *Helena*. The sizes of deletions and insertions are also very similar in all of the pseudogenes and *Helena*. Insertions are all small (from 1 to 5 bp, except for a single insertion of 35 bp in *D. sechellia Cecropin A2*), whereas some of the deletions are short ($<$10 bp) and some are much longer (up to 270 bp). Given the long right-hand tail of the distribution of deletion sizes it is difficult to judge whether deletion sizes in every pseudogene are drawn from the same distribution – the small numbers of deletions in each case afford little power to such a comparison. In addition the length of the pseudogenes themselves vary. Because longer deletions can be observed with higher probability (if at all) in longer pseudogenes, this introduces another difficulty in the analysis.

We can, however, get a rough sense that the distribution of deletion sizes in all of the pseudogenes and in DOA copies of *Helena* are similar. We can divide all deletions into two classes depending on whether they are smaller or larger than 10 bp. Focusing only on the *D. melanogaster Helena* dataset, approximately 50% of deletions fall into each class (34 out of 64 de-

Table 2. Comparison of the indel spectra in *Helena* non-LTR elements and *bona fide* pseudogenes in *D. melanogaster*

	Chromosome position	Ratio of deletions to nucleotide substitutions	Deletion sizes ($\leq$10 bp/$>$10 bp)	Ratio of deletions to insertions
Helena	Various; mostly in heterochromatin	0.13 (87/669)[a]	34/30	8.7 (87/10)
Larval cuticle protein ψ	Euchromatin, (65A)	0.07 (4/58) NS[b]	4/2	1.1 (6/5)[c*]
Swallow ψ[d]	Euchromatin, (5E)	0.10 (8/83) NS	4/4	(8/0)
Cecropin A2 ψ	Euchromatin (99E)	0.20 (2/25)[d] NS	3/1[d]	(4/0)[d]
			7/6[e]	6.5 (13/2) NS[e]
α Esterase 4a-ψ	Euchromatin (84D3-10)	0.15 (3/20) NS	0/3	(3/0)
All ψ-genes	Euchromatin	0.09 NS	15/15	4.3 (30/7) NS

[a]The numbers in brackets are the counts of different types of mutations.
[b]NS – not significant.
[c*] – Significant.
[d]Taking into account only the coding region alignment between *D. mauritiana CecA2*ψ and *D. melanogaster CecA2*.
[e]Taking into account all deletions and insertions in *D. mauritiana*, *D. sechellia*, and *D. simulans CecA2*ψ.

Table 3. Nuclear insertions of mitochondrial DNA in the sequenced portion of the *D. melanogaster* genome

Name of the clone[a]	Chromosome, map location	Length of homology	Region of homology[b]
AE003781	2L, 39E2	96 bp	2540–2636
AE003844	4, 102 A4-6	566 bp	1175–1773
AE003139	Unknown, possibly heterochromatin	151 bp	3400–3550

[a]Name of the clone from the BDGP project[50] that contains a particular numt.
[b]Relative to the sequence of the *D. melanogaster* mitochondrial genome (U37541).

letions are smaller than 10 bp). Deletions in the bona fide pseudogenes follow the same pattern – overall 15 out of 30 deletions are smaller than 10 bp.

These results show that most *Drosophila* pseudogenes lose DNA in a very similar way to the DOA copies of *Helena*. The high rate of DNA loss in *Drosophila* does not appear to depend on the exact sequence identity, on whether the sequence is transposable or not, or whether it resides in heterochromatin or euchromatin.

Nuclear insertions of mitochondrial DNA (numts) in Drosophila

Insertions of mitochondrial DNA into nuclear genomes have now been described in 83 different eukaryotes [5]. In Metazoa numts sequences are invariably nonfunctional [5] and thus represent another source of unconstrained DNA.

There are three recognizable insertions of mitochondrial DNA in the sequenced genome of *D. melanogaster* [5] (Table 3). They range in size from 96 to 566 bp and are derived from non-overlapping regions of mitochondrial DNA. Thus these sequences are unique in being recognizable, nonfunctional, and single-copy. Other recognizable, nonfunctional DNA is generally repetitive – transposable elements are often present in multiple copies and even single *bona fide* pseudogenes are similar in sequence to their functional paralogs. On the other hand, it is generally very difficult to identify any single copy DNA as nonfunctional. Given that DNA homology is known to exert multiple effects on the expression, chromatin structure, and even mutational processes in different organisms [34–38], it is entirely possible that repeated sequences would have a distinct pattern of indels compared to single-copy sequences.

Only one of the numts (from the clone AE003844) is sufficiently long to permit analysis of the indel pattern. Its alignment shows the presence of six deletions (three of 1 bp, 13 bp, 14 bp, and 27 bp) and no insertions. The ratio of six deletions to 0 insertions is statistically indistinguishable from that in *Helena* ($P = 0.9$, G-test with the Yates correction for con-

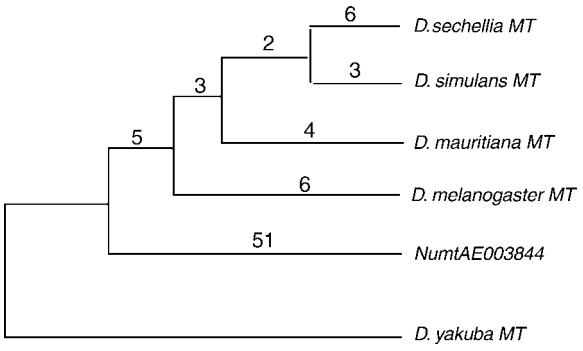

Figure 3. The best tree of mtDNA in the *D. melanogaster* subgroup and of *NumtAE003844*. I used the region of mtDNA homologous to NumtAE003844. Both maximum parsimony and maximum likelihood (HKY85) phylogeny reconstruction methods produced the same unique best tree.

tinuity). The distribution of deletion sizes is also very similar to the patterns observed in *Helena* and bona fide pseudogenes, with half of deletions smaller and half significantly longer than 10 bp.

To estimate the rate of deletions per point substitution I constructed a phylogeny of mtDNA and numtAE003844 in the *D. melanogaster* subgroup. Exhaustive searches using both maximum likelihood (HKY85) or maximum parsimony [39] resulted in the same unique best tree (Figure 3). Based on this tree, numtAE003844 inserted in the *Drosophila* nuclear genome approximately 5 MYA, shortly after the split of lineages leading to *D. melanogaster* and *D. yakuba* [40]. We can infer that since that time it has suffered 51 nucleotide substitutions (55 after the Jukes–Cantor correction for multiple hits), which corresponds to approximately 20×10^{-3} substitutions/nucleotide/Myr. This value is similar to the rate of synonymous substitution of the fast evolving *Drosophila* genes (16×10^{-3} substitutions/nucleotide/Myr) [24]. The ratio of 55 nucleotide substitutions to six deletions (0.11 deletions/substitution) is very similar to the relative rate of deletions in DOA copies of *Helena* ($P = 0.7$, G-test).

Overall, the indel spectrum in this numt is indistinguishable from that in DOA copies of *Helena* or in bona fide *Drosophila* pseudogenes. It appears that the high rate of DNA loss is shared by much of the complex, nonfunctional DNA in *Drosophila* regardless of its sequence or of whether it is unique or repetitive, transposable or non-transposable, and euchromatic or heterochromatic.

Mutational pattern in transcribed sequences

All nonfunctional *Drosophila* sequences discussed thus far are not transcribed. Could this be important? The answer is an unqualified yes in most organisms. Surprisingly, however, the answer maybe a no in *Drosophila*. There are many published reports on the molecular inter-relationship between transcription and repair (reviewed in [41]). It has been established in many organisms that both base excision repair and nucleotide excision repair operate much more efficiently in actively transcribed genes. Because differences in repair processes are likely to affect mutational biases, one would expect to find differences in mutational spectra in the transcribed and non-transcribed sequences.

Interestingly, however, transcription-coupled repair (TCR) is not likely to operate in *Drosophila*. Direct experiments failed to find any evidence for TCR in *Drosophila* [42, 43], and the analysis of the genome sequence failed to find homologues of the genes involved in TCR in other organisms [44]. Thus it is not clear whether mutational biases should differ between transcribed and non-transcribed sequences in *Drosophila*.

Nevertheless, such difference might still exist. In this respect, it is noteworthy that the imbalance between deletions and insertions is much less pronounced in introns (deletion-to-insertion ratio of 1.35) [45] than in non-transcribed pseudogenes, numts and defunct copies of *Helena*. Even though some of this difference is probably due to strong selection against deletions affecting splicing or spilling into the exons (S. Ptak, personal communication), or to weak selection for the increased recombination [45], it might also be partly due to different indel patterns in transcribed and non-transcribed DNA.

Natural selection and indel spectrum in Drosophila

The use of pseudogenes to study spontaneous mutation is usually justified by the fact that in the absence of coding function, all mutations in pseudogenes would have equal selective values and would be equally represented in the sample of pseudogene substitutions. However, this inference may be less straightforward than it may appear. First, some pseudogenes may be expressed at the level of RNA or protein for a period of time and such an expression could have a harmful effect. Mutations that inactivate transcription and/or translation of pseudogenes could be favored in such cases. Given that deletions (especially long ones) are more likely to have a severe effect on gene expression, such a selective effect would lead

to an overrepresentation of deletions among pseudogene substitutions [33]. Pesudogenes could also affect the expression of neighboring genes in a harmful way either directly, by carrying enhancer-like sequences, or indirectly, by changing the distances between regulatory elements or preventing communication among them. In such a case, again, deletions might be advantageous because they might be more likely to eliminate the harmful effects of pseudogenes on gene expression.

In addition, if the mere presence of the bulk pseudogene DNA is detrimental to the organism because it adds energetic costs and lengthens replication time, long deletions would be selectively favored, whereas longer insertions would be deleterious. If natural selection of either kind is sufficiently strong, then the patterns of substitutions observed in nonfunctional DNA will reflect both mutation rates and differential probabilities of persistence and fixation of deletions and insertions.

The possibility that deletions are overrepresented because they disrupt deleterious effects of pseudogenes on other genes (through expression or local effects) is highly unlikely to explain the high rate of observed DNA loss in *Drosophila*. It is clear that this effect would be highly specific to particular pseudogenes at the early stages of their deterioration. However, the similarity of indel spectra among different types of sequences of different ages and original function (or lack thereof, such as numtAE003844) argues that the right explanation of the high rate of DNA loss must apply to all *Drosophila* sequences equally.

The second source of selective bias – natural selection for a smaller genome size – would apply to all sequences in the genome and could, in principle, result in similarly biased profiles of observed indels in all nonfunctional sequences. Admittedly, indels representing approximately one ten millionth of the total genome size are unlikely to have very large selective coefficients. But in large enough populations, even small selective coefficients matter. Could it be that selection for smaller genome size is simply stronger or more efficient due to the larger population size [26] in *Drosophila* than in mammals, resulting in a higher rate and longer average size of deletions [46]?

There are several reasons to believe that our estimate of the pattern of indels in *Drosophila* is not appreciably biased by natural selection for a smaller genome size [47]. To the extent that indels have a selective impact because they change genome size, their selective coefficients must be proportional to their length. This follows simply from the fact that indels under consideration are exceedingly small compared to the genome size, and thus the fitness effect over such a small proportional change can be very well approximated by a linear function. The coefficient of linearity is, of course, unknown and can be large in principle. But the linearity itself puts constraints on the patterns of bias we expect to see under this selective scenario.

In particular, if deletions of ~25 bp in length are beneficial to a degree sufficient to affect their representation ($N_e s > 1$), the original insertions of 1.3 kb long insertions of *Helena* elements present in our dataset should be extremely deleterious ($N_e s < -50$). Such a strong negative selection is inconsistent with the long persistence of many *Helena* elements (on average 26 million generations) and with an observation of fixation of one of *Helena* elements in the *Drosophila* data. In particular, fixation of an element with such a strong negative selective coefficient would imply an absurdly high rate of transposition ($\sim 10^{12}$ transpositions per individual per generation).

Another observation inconsistent with the linear response of selection to an indel size is the fact that only deletions larger than 5 bp are significantly more frequent in *Drosophila Helena* elements than in mammalian pseudogenes. In fact, deletions of 3–5 bp are found in equal frequencies in *Drosophila* and mammals (G-test, $P = 0.75$), whereas deletions of 6–8 bp are 25-fold more frequent in *Drosophila* (G-test, $P = 7 \times 10^{-5}$). This observation is inconsistent with a linear increase of selective coefficients with the deletion length [47]. Selection for a smaller genome size in *Drosophila* should not result in such a sharp difference ($P = 0.015$) [47], whereas selection for a larger genome in mammals should produce a much lower proportion of deletions longer than 11 bp ($P = 1 \times 10^{-11}$).

It is clear that the difference in our estimates of the rate of DNA loss through small deletions between *Drosophila* and mammals is largely or even exclusively due to a difference in the mutational pattern itself and not to a differential effect of selection on segregating indels in these two taxa. However, even though the above considerations do show that natural selection is not strong enough to significantly bias our indel data, they do not imply in any way that natural selection cannot be strong enough to affect larger genome size variants in *Drosophila*.

Conclusions

Recent studies of indel rates and patterns strongly suggest that variation in indel spectra may be an important factor in genome size evolution. For example, *Drosophila*, which has a compact genome with little extra DNA such as pseudogenes, spontaneously loses DNA at a much higher rate than *Podisma* grasshoppers [9], Hawaiian crickets (Genus *Laupala*) [8], or mammals [6, 7, 20–23] – all of which have much larger genomes, with a higher proportion of noncoding DNA. In addition *C. elegans* has a small genome and accordingly its pseudogenes appear to have a high rate of relatively large deletions [48]. In plants, small-genome *Arabidopsis* tends to leave larger deletions after the repair of double strand breaks than large-genome tobacco plants [49]. All of these data suggest not only that indel spectra vary a great deal among taxa, but also that these differences probably contribute to the large scale differences in genome size.

Most of these studies rely on different sources of nonfunctional DNA to infer indel spectra characteristic of different genomes. The evidence summarized in this paper demonstrate that at least in the case of *Drosophila*, where most of the data are available, estimates derived from different kinds of nonfunctional DNA agree with one another and that these estimates are essentially free of selective biases. These results should be seen both as an encouragement and as a warning. They should encourage us to continue the study of deletion biases in different taxa using available nonfunctional DNA such as pseudogenes, DOA non-LTR elements and other defunct transposable elements, and numts. However, even in *Drosophila* we cannot currently determine whether transcribed sequences suffer a different indel pattern than non-transcribed DNA. In other organisms the situation is often worse, where we have estimates derived from only a single kind of nonfunctional DNA when possible effects of natural selection on the observed data cannot be ascertained. We should be very careful in our interpretations of such data, for even though any single source of (untranscribed) nonfunctional DNA gives the same answer in *Drosophila*, the same may not be true for other taxa or kinds of sequences.

References

1. Yampolsky L. Y. and Stoltzfus A.: Evol. Dev. 3 (2001): 73–83.
2. Stoltzfus A.: J. Mol. Evol. 49 (1999): 169–181.
3. Li W. H., Wu C. I. and Luo C. C.: J. Mol. Evol. 21 (1984): 58–71.
4. Gojobori T., Li W. H. and Graur D.: J. Mol. Evol. 18 (1982): 360–369.
5. Bensasson D., Zhang D.-X., Hartl D. L. and Hewitt G. M.: Trends Ecol. Evol. 16 (2001): 314–321.
6. Petrov D. A., Lozovskaya E. R. and Hartl D. L.: Nature 384 (1996): 346–349.
7. Petrov D. A. and Hartl D. L.: Mol. Biol. Evol. 15 (1998): 293–302.
8. Petrov D. A. et al.: Science 287 (2000): 1060–1062.
9. Bensasson D. et al.: Mol. Biol. Evol. 18 (2001): 246–253.
10. Petrov D. A.: Trends Genet. 17 (2001): 23–28.
11. Petrov D. A. and Hartl D. L.: Gene 205 (1997): 279–289.
12. Hardies S. C. et al.: Mol. Biol. Evol. 3 (1986): 109–125.
13. Luan D. D., Korman M. H., Jacubczak J. L. and Eickbush T. H.: Cell 72 (1993): 595–605.
14. Lathe W. C., Burke W. D., Eickbush D. G. and Eickbush T. H.: Mol. Biol. Evol. 12 (1995): 1094–1105.
15. Burke W. D., Malik H. S., Lathe W. C., 3rd and Eickbush T. H.: Nature 392 (1998): 141–142.
16. Weiner A. M., Deininger P. L. and Efstratiadis A.: Annu. Rev. Biochem. 55 (1986): 631–661.
17. Malik H. S., Burke W. D. and Eickbush T. H.: Mol. Biol. Evol. 16 (1999): 793–805.
18. Hutchison III C. A. et al.: In: Berg D. E. and Howe M. M. (eds), Mobile DNA. American Society for Microbiology, 1989, pp. 593–617.
19. Petrov D. A., Schutzman J. L., Hartl D. L. and Lozovskaya E. R.: Proc. Natl. Acad. Sci. USA 92 (1995): 8050–8054.
20. Graur D., Shuali Y. and Li W. H.: J. Mol. Evol. 28 (1989): 279–285.
21. Gu X. and Li W. -H.: J. Mol. Evol. 40 (1995): 464–473.
22. Ophir R. and Graur D.: Gene 205 (1997): 191–202.
23. Robertson H. M. and Martos R.: Gene 205 (1997): 219–228.
24. Sharp P. M. and Li W. -H.: J. Mol. Evol. 28 (1989): 398–402.
25. Bensasson D., Zhang D. X. and Hewitt G. M.: Mol. Biol. Evol. 17 (2000): 406–415.
26. Jensen S. and Heidmann T.: EMBO J. 10 (1991): 1927–1937.
27. Pelisson A., Finnegan D. J. and Bucheton A.: Proc. Natl. Acad. Sci. USA 88 (1991): 4907–4910.
28. Lozovskaya E. R., Nurminsky D. I., Petrov D. A. and Hartl D. L.: Genes Genet. Syst. 74 (1999): 201–207.
29. Lozovskaya E. R., Scheinker V. S. and Evgen'ev M. B.: Genetics 126 (1990): 619–623.
30. Petrov D. A., Chao Y.-C., Stephenson E. C. and Hartl D. L.: Mol. Biol. Evol. 15 (1998): 1562–1567.
31. Pritchard J. K. and Schaeffer S. W.: Genetics 147 (1997): 199–208.
32. Ramos-Onsins S. and Aguade M.: Genetics 150 (1998): 157–171.
33. Robin G. C., Russell R. J., Cutler D. J. and Oakeshott J. G.: Mol. Biol. Evol. 17 (2000): 563–575.
34. Selker E. U.: Trends Genet. 13 (1997): 296–301.
35. Birchler J. A., Pal-Bhadra M. and Bhadra U.: Nat. Genet. 21 (1999): 148–149.
36. Pal-Bhadra M., Bhadra U. and Birchler J. A.: Cell 90 (1997): 479–490.
37. Yoder J. A., Walsh C. P. and Bestor T. H.: Trends Genet. 13 (1997): 335–340.
38. Henikoff S. and Matzke M. A.: Trends Genet. 13 (1997): 293–295.

39. Swofford D. L.: PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4, Sinauer Associatesx, 2001.
40. Russo C. A. M., Takezaki N. and Nei M.: Mol. Biol. Evol. 12 (1995): 391–404.
41. de Laat W. L., Jaspers N. G. and Hoeijmakers J. H.: Genes Dev 13 (1999): 768–785.
42. de Cock J. G. et al.: Nucl. Acids Res. 20 (1992): 4789–4793.
43. van der Helm P. J., Klink E. C., Lohman P. H. and Eeken J. C.: Mutat. Res. 383 (1997): 113–124.
44. Sekelsky J. J., Brodsky M. H. and Burtis K. C.: J. Cell. Biol. 150 (2000): F31–F36.
45. Comeron J. M. and Kreitman M.: Genetics 156 (2000): 1175–1190.
46. Charlesworth B.: Nature 384 (1996): 315–316.
47. Petrov D. A. and Hartl D. L.: J. Hered. 91 (2000): 221–227.
48. Robertson H. M.: Genome Res. 10 (2000): 192–203.
49. Kirik A., Salomon S. and Puchta H.: Embo. J. 19 (2000): 5562–5566.
50. Adams M. D. et al.: Science 287 (2000): 2185–2195.