Pseudogene Evolution and Natural Selection for a Compact Genome

D. A. Petrov and D. L. Hartl

Pseudogenes are nonfunctional copies of protein-coding genes that are presumed to evolve without selective constraints on their coding function. They are of considerable utility in evolutionary genetics because, in the absence of selection, different types of mutations in pseudogenes should have equal probabilities of fixation. This theoretical inference justifies the estimation of patterns of spontaneous mutation from the analysis of patterns of substitutions in pseudogenes. Although it is possible to test whether pseudogene sequences evolve without constraints for their protein-coding function, it is much more difficult to ascertain whether pseudogenes may affect fitness in ways unrelated to their nucleotide sequence. Consider the possibility that a pseudogene affects fitness merely by increasing genome size. If a larger genome is deleterious—for example, because of increased energetic costs associated with genome replication and maintenance—then deletions, which decrease the length of a pseudogene, should be selectively advantageous relative to insertions or nucleotide substitutions. In this article we examine the implications of selection for genome size relative to small (1–400 bp) deletions, in light of empirical evidence pertaining to the size distribution of deletions observed in *Drosophila* and mammalian pseudogenes. There is a large difference in the deletion spectra between these organisms. We argue that this difference cannot easily be attributed to selection for overall genome size, since the magnitude of selection is unlikely to be strong enough to significantly affect the probability of fixation of small deletions in *Drosophila.*

The Darwinian theory of evolution treats heritable variation as random and undirected. Mutation merely supplies the raw material for natural selection, which is seen as the directional force that drives evolution toward greater adaptation. But "random" is a tricky word. Although mutational variation is not generally biased toward any particular adaptation, this does not imply that mutational variation is random in other respects. Some variants arise by mutation more frequently than others, independently of natural selection, and these also affect the ultimate course of evolution since the probability of fixation under mutation–selection–drift is a function of the mutation rate.

The possible existence of biases in spontaneous mutation, and the importance of assessing these biases quantitatively, have been recognized for a long time (Beale and Lehmann 1965; Li et al. 1984, 1985; Zuckerkandl et al. 1971). However, empirical studies have proven difficult. Because spontaneous mutation is generally too rare to be investigated directly in laboratory studies, and because laboratory studies are almost inevitably subject to experimental bias according to the methods by which mutations are detected, attention has been devoted primarily to making inferences about patterns of spontaneous mutation from analyses of the observed patterns of nucleotide substitution in functional genes. Although this approach is very powerful, its reliability remains questionable. The main problem is that patterns of substitution in functional genes are affected not only by the relative rates at which different mutations occur spontaneously, but also—and often decisively—by natural selection. Some types of mutation may be, on average, more deleterious than others, and as a result, such mutations will be underrepresented in any observed sample.

The problem of bias is most acute for substitutions in coding sequences. In the exons of protein-coding genes, for example, insertions and deletions (indels) are observed much less frequently than point substitutions (simple nucleotide substitu-

From the Harvard University Society of Fellows (Petrov) and the Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138. Address correspondence to Daniel L. Hartl at the address above or e-mail: dhartl@oeb.harvard.edu. This paper was delivered at a symposium entitled "Genetic Diversity and Evolution" sponsored by the American Genetic Association at the Pennsylvania State University, University Park, PA, USA, June 12–13, 1999.

 91:221–227

tions). This observation tells us essentially nothing about the relative rates of spontaneous indel formation versus point substitution, since indels in exons are almost certain to be severely deleterious. In this example, the possible effects of mutational bias on the fixation of mutations are hopelessly confounded with the effects of selection.

An alternative to dealing with the biases inherent in studying functional genes is to study their nonfunctional counterparts—pseudogenes—instead. Because pseudogenes are presumed to evolve without functional constraints, it can be argued that all types of mutation should have an equal chance of fixation (or persistence in populations), and thus the pattern of substitutions in pseudogenes should be congruent with the pattern of mutations (Graur et al. 1989; Li et al. 1981, 1984). This approach has been widely used to estimate patterns of mutation in mammals (Graur et al. 1989; Li et al. 1981, 1984). More recently we invoked essentially the same logic to estimate patterns of mutation in *Drosophila* using immobile, nonfunctional, "dead-on-arrival" (DOA) copies of non-LTR (long terminal repeat) retrotransposable elements in lieu of conventional pseudogenes (Petrov et al. 1996; Petrov and Hartl 1998).

How confident can one be that a sequence designated as a pseudogene does, in fact, have no coding function? The current standard for classifying a sequence as a pseudogene is to show either that it is not transcribed and translated, or that it lacks a complete open reading frame (ORF). Lack of coding function can also be argued on molecular evolutionary grounds if the pattern of nucleotide substitution in the sequence shows no evidence of functional constraints on the coding capacity. In practice, this usually means that the rate of synonymous and nonsynonymous nucleotide substitutions are identical ($Ks/Ka \approx 1$), and that the ORF in the sequence is disrupted by stop codons, deletions, and/or insertions.

Current criteria for pseudogenes address only the protein-coding function of functional genes. But some pseudogene sequences may have position effects on the expression of nearby genes, or they may be mutagenic through homologous DNA interactions with their functional counterparts (Wu and Morris 1999). When position effects occur, most of them would be expected to be deleterious, but a few may be beneficial. In cases where pseudogenes exert detrimental effects on the expression of nearby genes, more drastic mutations in the pseudogene may be selectively advantageous, as they are more likely to disrupt the deleterious activity of the pseudogene. The reverse will be true if the activity of a pseudogene is advantageous. The problem is that generally one does not know whether a pseudogene has any noncoding phenotypic effect and whether the effect is deleterious or advantageous.

Besides having position-specific or sequence-specific effects on the expression of genes, pseudogenes may exert more global effects through their aggregate effects on genome size. If a compact genome is favored by selection, then it follows that deletions occurring in pseudogenes should be slightly beneficial and insertions slightly deleterious, even though point substitutions may be selectively neutral. On the other hand, if a large genome is favored by selection, then deletions will be slightly deleterious and insertions slightly beneficial. We emphasize the word "slightly." There's the rub, for the effectiveness of selection on any mutant allele is determined by the magnitude of $N_e s$, where N_e is the effective population number and s is the selection coefficient. If $N_e s$ is large enough, then the probabilities of fixation will be skewed from the spontaneous mutation frequencies, and so inferences about spontaneous mutation from the analysis of pseudogene substitutions become problematic.

In this article we consider the model of global selection for genome size and consider its implications for the probability of fixation of small deletions and insertions in individual copies of pseudogenes scattered throughout the genome. The central issue is the likely magnitude of $N_e s$ for small indels, which we discuss in light of the observed size distributions of indels. Based on a number of considerations we conclude that natural selection for total genome size must have a negligible effect on small (<400 bp) deletions and insertions. Hence we affirm an earlier suggestion that the insertion/deletion spectra found in pseudogenes or in DOA copies of non-LTR elements may generally afford nearly unbiased estimates of the insertion/deletion spectra among spontaneous mutations (Petrov et al. 1996).

Background and Theoretical Considerations

The data we examine are estimates of the insertion/deletion spectra in mammals based on an analysis of pseudogenes (Graur et al. 1989) and in *Drosophila* based on DOA copies of non-LTR retrotransposable elements (Petrov et al. 1996, Petrov and Hartl 1998). In these studies genome size varies from 1.6×10^8 bp in *Drosophila* (a relatively compact genome) to 3.0×10^9 bp in humans (a relatively large genome). The deletion/insertion sizes ranged from 1 bp to approximately 500 bp and were limited in part by the length of the pseudogene sequences studied (~0.4–2 kb). Although the estimated values are specific to the organisms examined, the implications from them are more general in that the sizes of the indels examined are much smaller than the total genome size.

The Shape of the Selection Curve for Small Deletions and Insertions

We consider a model of global selection for genome size in which noncoding DNA sequences are subjected to selection only insofar as they affect genome size. In this model we ignore any selective effects due to local position effects on gene expression, and any other sequence-specific or position-dependent mechanisms of selection.

Consider a newly formed indel (deletion or insertion) of a length ΔG bp. The presence of such an indel changes the genome size by the value of its length, namely, ΔG bp. On average, the proportionate change in the genome size due to such an indel will be $\Delta G/G_0$, where G_0 bp is the average genome size in the population. If $\Delta G/G_0$ is small, then the selection coefficient, $s(\Delta G/G_0)$, associated with the indel in question, owing only to its effect on genome size, can be found by expanding $s(\Delta G/G_0)$ around $\Delta G/G_0 = 0$ in a Taylor series,

$$\begin{aligned} s(\Delta G/G_0) &= s(0) + s'(0)[\Delta G/G_0] \\ &= s'(0)[\Delta G/G_0] \end{aligned} \qquad (1)$$

where $s(0) = 0$ because the fitnesses are measured relative to the current average genome size in the population, and $s'(0)$ is the slope of the selection coefficient function evaluated at the current average genome size. If $s'(0) > 0$ then a genome larger than the current average is locally favored, if $s'(0) < 0$ then a genome more compact than the current average is locally favored, and if $s'(0) = 0$ then the current average is a local optimum of genome size (the alternative possibility that G_0 defines a local fitness minimum seems too remote to consider).

It is unlikely that the current average genome size represents the optimal value,

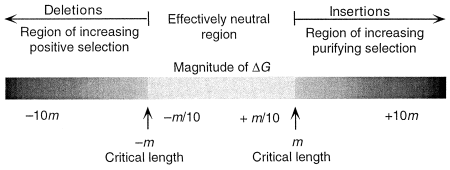

Figure 1. The "critical length" of an indel is here defined as the length for which $|4N_e s| \approx 1$. Indels with lengths smaller than the critical length are nearly neutral. When there is genome-wide selection for a compact genome size, indels larger or smaller are subjected to a selection coefficient proportional to their size, positive if the indel is a deletion and negative if it is an insertion. For $|4N_e s| \gg 1$, the ultimate fate of an indel is determined almost exclusively by selection.

because there is undoubtedly some mutational load, due to multiple sources, of indels being formed continuously. This means that $s'(0) \neq 0$, and equation (1) therefore implies that, for indels that are small relative to the current average genome size, the selection coefficient should be linear in the magnitude of the change in genome size. For the sizes of indels observed in DOA retrotransposons in *Drosophila* and pseudogene mammals (1–1000 bp), relative to the total genome size (10^8–10^9 bp), the values of $\Delta G/G_0$ are in the range 10^{-5}–10^{-9}, so the assumption in equation (1) that $\Delta G/G_0$ is small is certainly justified.

Consequences of Linear Fitness Response to Small Changes in Genome Size

The ultimate fate of an indel of a particular size $\Delta G \ll G_0$ is determined by the product of the selection coefficient s, $s(\Delta G/G_0)$, and the reciprocal of four times the effective population number, $1/4N_e$. In particular, the magnitude of $|4N_e s|$ determines whether the selection-drift process leading ultimately to fixation or loss is essentially stochastic ($|4N_e s| \ll 1$), essentially deterministic ($|4N_e s| \gg 1$), or subject to nonnegligible effects from both selection and random genetic drift ($|4N_e s| \approx 1$). In Figure 1 we designate the critical length, *m,* as the size of an indel, ΔG, for which $|4N_e s| = 1$. Because of the linearity of the selection coefficient [equation (1)], it follows that the ultimate fate of any indel that is smaller than *m* by a factor of 10 or more will be essentially determined by random genetic drift ($|4N_e s| \leq 0.1$), whereas the ultimate fate of any indel that is larger than *m* by a factor of 10 or more will be essentially determined by selection ($|4N_e s| \geq 10$) (Figure 1).

Unfortunately the critical indel length *m* is unknown for any organism. We can nevertheless make some important inferences

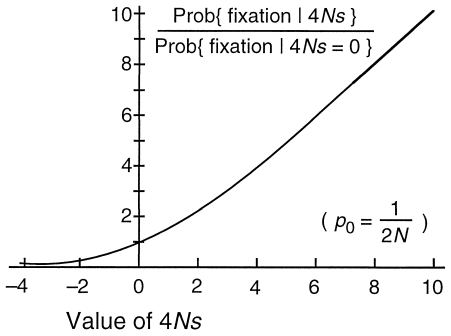

Figure 2. Probability of ultimate fixation of a newly arising indel with a value of $4N_e s$ given along the abscissa, relative to a newly arising neutral mutation with $4N_e s = 0$.

by identifying a range of indel sizes over which the extremes give evidence of being subjected to large selective forces of opposite sign, for then the critical length must be within this range. For example, assuming a uniform distribution of indel sizes, if one could show that deletions of size 10 bp were present at much lower frequencies in a population sample than those of 1000 bp, and that deletions of 1000 bp were present at much higher frequencies than insertions of 1000 bp, then one could infer that the critical length must be somewhere in the range 10–1000 bp and, moreover, that deletions are selectively favored. The same inference would follow from the finding of drastically different fixation probabilities, or persistence times, for deletions and insertions spanning the 10–1000 bp interval. The key point is that, in view of the linearity of the selection coefficient as a function of indel size, so long as the indels are small, then the interval over which selective difference should be detectable at the extremes includes the critical length.

Figure 2 shows the probability of fixation of a newly arising indel with some value of $4Ns$, relative to the probability of fixation of a newly arising neutral allele ($4Ns = 0$), assuming $N_e = N$. This ratio of probabilities is given explicitly by $4Ns/[1 - \exp(-4Ns)]$, and the range of values of $4Ns$ has been chosen so that the ratio of fixation probabilities ranges from approximately 1/10 (actually 0.075) for $4Ns = -4$ to 1/10 for $4Ns = +10$. This curve implies that an indel favored by selection must be 10 times longer than the critical value in order to have about a sixfold greater probability of fixation than an indel matching the critical value. An increase of sixfold in the probability of fixation is roughly what one might expect to be detectable in datasets of the size of those presently available. Hence if the critical value were small (say, 10 bp), then there would be a noticeable effect of selection for indels larger than 100 bp, whereas if the critical value were large (for example, 100 kb), then only very large indels on the order of 1 Mb would have a greatly increased chance of fixation due to positive selection.

Empirical Evidence

In the rest of this article we examine empirical distributions of sizes of deletions and insertions in DNA sequences in *Drosophila* and mammals that are unconstrained with respect to their protein-coding function. We concentrate primarily on the possibility that the difference in the observed distributions, which is skewed toward more deletions and longer deletions in *Drosophila,* is due to stronger natural selection for a compact genome in *Drosophila* as compared to that in mammals. We conclude that, in *Drosophila,* the critical indel length at which $|4N_e s| = 1$ is larger (and possibly much larger) than the size of indels we observe. Consequently the ultimate fate of individual indels in this size range should be determined largely, if not exclusively, by random genetic drift—even if *Drosophila* is subjected to a genome-wide selective force favoring a compact genome.

Defunct Transposable Elements as Pseudogene Surrogates

One of the most enigmatic differences in genome organization of *Drosophila* and mammals is the drastically different frequencies of pseudogenes. Mammalian genes often have tens or even hundreds of pseudogene counterparts (Weiner et al. 1986), whereas pseudogenes in *Drosophila* are exceedingly rare (Jeffs and Ashburner 1991). The paucity of pseudogenes in *Drosophila* has hampered studies of DNA sequences that are unconstrained with respect to their protein-coding function in this organism and has precluded comparative studies of pseudogene evolution outside of mammals.

We have proposed that the molecular evolutionary information that can be gleaned from bona fide pseudogenes (nonfunctional copies of single-copy functional genes) can be obtained in *Drosophila* and other organisms from the study of defunct, DOA copies of non-LTR retrotransposable elements (Petrov et al. 1996; Petrov and Hartl 1997, 1998). Here we use *defunct* in the proper dictionary sense of "no longer living, dead," because non-LTR

elements commonly create truncated, immobile, and nonfunctional copies that are "virtual" pseudogenes in the sense that they are unconstrained in their protein-coding function. We have shown how phylogenetic analyses of multiple, independently transposed copies of a non-LTR element can be used to separate the constrained evolution of lineages of transpositionally active elements from the unconstrained, pseudogene-like evolution of lineages of defunct elements (Petrov et al. 1996; Petrov and Hartl 1997, 1998). Because non-LTR elements are virtually ubiquitous in eukaryotes, and can easily be cloned or amplified with universal primers, the approach using defunct non-LTR elements should provide a general means to study unconstrained DNA in practically any eukaryote.

Pseudogene Evolution in *Drosophila* and Mammals

Application of this approach allowed the first estimate of the pattern of spontaneous substitutions in *Drosophila* pseudogenes, revealing features of evolutionary stability as well as features of extreme variation. The pattern of spontaneous, simple nucleotide substitutions (point substitutions) in *Drosophila* pseudogenes proved to be surprisingly similar to that in mammals, despite the very long evolutionary distance separating these animals. In fact, excluding the much higher rate of G::C to A::T transition in mammals (probably attributable to methylation of cytosines in mammals but not in *Drosophila*), the patterns are statistically indistinguishable (Petrov and Hartl 1999). The same is true for insertions: in both organisms pseudogenes accumulate on average 1–1.5 small insertions (2–3 bp in length) per 100 point substitutions. On the other hand, the pattern of deletions is very different in *Drosophila* than in mammals. In *Drosophila,* deletions in defunct retroelements are 2.5 times more frequent, and on average 7 times longer, than in mammals. These differences alone imply almost a 20-fold higher rate of loss of unconstrained DNA in *Drosophila* (Petrov et al. 1996; Petrov and Hartl 1998).

Differences in deletion spectra

Surprisingly, the difference in the size of deletions between *Drosophila* and mammals is due exclusively to a much higher incidence of deletions exceeding 5 bp in *Drosophila* (Figure 3). In particular, the rates of deletions of 1–5 bp are indistinguishable in *Drosophila* and mammals (G test, $P = .96$), whereas deletions of 6–15 bp are sixfold more common in *Drosophila* (G test, $P = 2 \times 10^{-5}$), and deletions of 16–30 bp are fourfold more frequent in *Drosophila* (G test, $P = .01$).

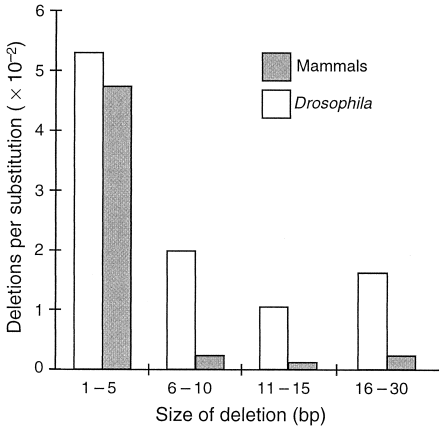

Figure 3. Size distributions observed in pseudogenes in mammals and in defunct retroelements in *Drosophila.* The numbers in the size range 1–5 bp are not significantly different.

The difference in the deletion spectra between *Drosophila* and mammals may be due to differences in spontaneous deletion formation, on the one hand, or to differences in the strength and direction of natural selection, on the other hand. If it is due to differences in spontaneous deletion formation, then the finding implies only that deletions larger than 5 bp are generated at a much greater frequency in *Drosophila* than in mammals. If one rejects this hypothesis, and instead invokes genome-wide selection for a compact genome to explain the difference, then the implication is that deletions larger than 5 bp have a greater probability of fixation and a longer persistence in *Drosophila* than in mammals. To explain why selection affects deletions larger than 5 bp in *Drosophila* more than in mammals, one would have to argue that the selection for a compact genome is either stronger in *Drosophila,* for whatever reasons, or that it is more effective in *Drosophila* because of a larger effective population number.

Can Selection Alone Explain the Difference in Deletion Spectra Between Mammals and *Drosophila*?

The null hypothesis under consideration is that the underlying distribution of spontaneous deletions is the same in *Drosophila* as in mammals, and that the observed differences in the deletion spectra result solely from genome-wide selection for a compact genome in *Drosophila* (Charlesworth 1996). If this were true, then it is unclear how selection for a compact genome could create such a sharp difference in probability of fixation over the small range of deletion sizes shown in Figure 3. There is a very sharp discontinuity at 5 bp. Deletions of 3–5 bp are found at equal frequency in *Drosophila* and mammals (G test, $P = .75$), whereas deletions of 6–8 bp are 25-fold more frequent in *Drosophila* (G test, $P = 7 \times 10^{-5}$).

To clarify the implications of Figure 3 relative to the null hypothesis, assume that all of the indels observed in mammalian pseudogenes and in defunct retroelements in *Drosophila* are fixed (not polymorphic) in the population. This is a conservative assumption with regard to the null hypothesis, because to assume that a deletion has become fixed in the population is more favorable for positive selection than to assume that it is a segregating polymorphism. Now consider the effect of positive selection for a compact genome, given that, over the range of indel sizes in Figure 3, the selection coefficient must be linear in indel size. Because there is no difference in the fixation of 3–5 bp deletions, we have to assume that $4N_es$ is between 0 and 1 for deletions of 3–5 bp. Therefore, because of linearity, the value of $4N_es$ for deletions of 6–8 bp should be no more than 3 (and smaller on the average). The increase in the fixation probability of mutations with $4N_es$ of 3, compared with neutral mutations, is 3.16 (Figure 2). But this is very much smaller than the 25-fold elevation actually observed (likelihood ratio test, $P = .015$).

An alternative scenario is to suppose that longer deletions are deleterious in mammals owing to genome-wide selection for a larger genome, whereas deletions in *Drosophila* are not subjected to selection. If we suppose that $4N_es = -1$ for mammalian deletions of 3–5 bp, then we infer $4N_es = -3$ for deletions of 6–8 bp. Based on these values we would predict about a sixfold deficit in the observed frequency of deletions of 6–8 bp in mammals, which is smaller than the deficit that we observed but not significantly so (likelihood ratio test, $P = .13$). However, for deletions in the size range 11–15 bp, we would expect $4N_es \leq -5$, which predicts at least a 30-fold deficit in this range of sizes, whereas the observed reduction is only sevenfold. For deletions in the size range of 16–30 bp, we would expect $4N_es \leq -9$, which

predicts at least a 900-fold deficit in this range of sizes, while the observed reduction is only fourfold. An omnibus likelihood ratio test demonstrates that such a scenario is extremely unlikely ($P = 1 \times 10^{-11}$).

No matter what scenario of genome-wide selection might be invoked, unless the underlying size distributions of spontaneous deletions are different in mammals and *Drosophila,* the finding of equal frequencies of deletions of 3–5 bp implies that $|4N_e s| \approx 1$ for this range of sizes, and then the implications of equation (1) for linearity of the selection coefficients is inescapable. It follows that at least part of the difference in the deletion spectra must be attributed to real differences in the size distribution of spontaneous deletions.

Can the Critical Deletion Size in *Drosophila* be as Small as Those Observed?

So far we conclude that genome-wide selection alone cannot explain the discrepancy in the deletion spectra between *Drosophila* and mammals. But genome-wide selection could still account for part of the difference, and this is the possibility that we shall now consider.

Each defunct retroelement begins its evolutionary lineage as a unique transposition event—a single insertion somewhere in the genome—that occurs in a single individual and therefore has an initial frequency in the entire population of $1/2N$. In other words, each defunct retroelement begins as an insertion/deletion polymorphism in the population, in which the insertion has a frequency of $1/2N$ and the "deletion" (absence of the defunct element) has a frequency of $1 - 1/(2N)$. If there were genome-wide selection for a compact genome, then this selection must also operate on the insertion/deletion polymorphism associated with the creation of each defunct element. This line of reasoning makes it difficult to fathom why defunct elements should persist in the population long enough to accumulate point substitutions and indels within themselves, let alone to become fixed. One could argue that certain regions of the genome may be more permissive to the fixation of defunct elements, either because they are somehow protected from genome-wide selection for a compact genome, or because they are in a region of high background selection (Charlesworth et al. 1995), and so can become fixed by chance, in spite of being deleterious. Such conditions might be found in centromeric heterochromatin. However, if defunct elements are fixed because they happen to reside in regions protected from selection for genome size, then how could selection for genome size operate with respect to indel mutations that occur within these very elements? And if they reside in regions in which selection is ineffective because of high background selection, then again how could the selection be more effective with respect to indel mutations that occur within the selfsame elements?

Although we have made no attempt to estimate what fraction of the defunct retroelements in the *Drosophila* dataset are fixed in any particular species, we have noted a 1300 bp element that was fixed in the common ancestor of the *D. simulans, D. mauritiana,* and *D. sechellia* clade (Petrov and Hartl 1998). Consider this 1300-bp element in relation to the average size of the deletions observed in all defunct elements, which is 25 bp. If we assume that $N_e s = 1$ for a deletion of 25 bp (assuming any smaller value makes the following arguments even stronger), then $N_e s$ for a 1300 bp deletion would be -52. The probability of fixation of the initial 1300-bp polymorphism, given $N_e = 10^6$ (Akashi 1997) and $N = N_e$, is about 7×10^{-28}. To get a sense of how small this probability is, imagine that we were in a position to observe every single fixation of a 1300 bp element over 10^9 generations ($\sim$10 million years for *Drosophila*) in a population of 10^7 individuals. In order to observe, on average, one such fixation, the rate of transposition would have to be on the order of 10^{12} per genome per individual. This is clearly an absurdly high transposition rate, even under the most congenial assumptions.

Even if the critical value at which $N_e s = 1$ were a deletion size of 50, then $N_e s$ for a 1300 bp polymorphism would be -26 and the probability of fixation would be 7×10^{-17}. Granted that this is just one example: perhaps this particular defunct element does reside in a region of relaxed selection, or is even beneficial owing to a local position effect. On the other hand, this particular element is typical: the average rate of deletions and the average size of deletions are indistinguishable from those of other defunct elements (G test for rate, $P = 0.42$; Wilcoxon test for size, $P = 0.87$).

Even if special circumstances account for the fixed element of 1300 bp, these circumstances are not expected to apply broadly to other defunct elements. Under the hypothesis of genome-wide selection for a compact genome, owing to size alone, we expect $|4N_e s|$ for each initially polymorphic insertion to be much greater than the $|4N_e s|$ for a 25 bp deletion. Such deleterious alleles are not expected to achieve high frequencies in the population or, equivalently, to persist in the population for long periods of time. The persistence times of defunct elements can be estimated from the number of apomorphic (terminal-branch) nucleotide substitutions per site within these elements (Petrov et al. 1996; Petrov and Hartl 1997, 1998). If we assume a constant rate of neutral point mutations of 15×10^{-3} per million years (Sharp and Li 1989), the average age of defunct elements in the *D. melanogaster* dataset is 1.33 ± 0.292 million years (range 0.1–4.8 million years), which corresponds to approximately 26 million generations assuming about 20 generations per year. This is a very long persistence time for large insertions, which under the null hypothesis should be very deleterious relative to the average size of indels. By way of comparison, for new neutral alleles in a population of size $N_e = N = 10^6$, the average time to loss of alleles destined to be lost is about $2 \ln(2N) = 29$ generations. We might argue that the persistent elements are not destined to be lost, but rather are fixed or destined to be fixed—but this merely throws us back on the other horn of the dilemma of specifying a mechanism by which large numbers of deleterious insertions can become fixed.

Without invoking an ad hoc assumption that each defunct element in our study was positively selected for some unknown favorable local effect, we must conclude that the critical indel length in *Drosophila* is larger than the size of most of the indels observed. Furthermore, the long persistence times and high probabilities of fixation of defunct elements suggest that the critical value may be much larger.

Population Persistence and the Length of Deletions

Deletions in the *D. melanogaster* dataset range from 1 to 432 bp. Under the null hypothesis of genome-wide selection for a compact genome, if the smallest selection coefficient yields $4N_e s \approx 1$, then assuming linearity [equation (1)] the largest should yield $4N_e s \approx 100$. Since the persistence times of mutations that differ by such a large range of selective values are also very different, we should see some indication that defunct elements with large de-

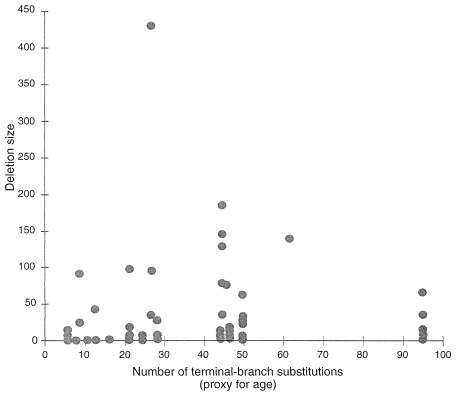

Figure 4. Deletion size (ordinate) as a function of age of the defunct element in which the deletion is found (abscissa). There is no significant correlation.

letions persist longer than defunct elements with small deletions.

We can use the number of accumulated point substitutions to estimate the age of a particular defunct element assuming a nucleotide substitution rate of 15×10^{-3} per million years (Sharp and Li 1989). The ages of defunct elements in the *D. melanogaster* dataset range from 0.1 to 4.8 million years, and there is a very good correlation between the number of deletions and the age of individual elements (Petrov and Hartl 1998). The simplest interpretation of these data is that deletions have had more time to accumulate in older elements and thus, on average, deletions found in older elements should be older than deletions found in younger elements.

If there is genome-wide selection for a compact genome, then longer deletions, being more strongly favored, should persist longer than small deletions, and the effect should be detectable, since it involves values of $4N_e s$ ranging over a factor of 100. By "persistence" we do not necessarily mean persistence as polymorphisms, merely long-term persistence in the genome, even if fixed. This possibility would yield a positive correlation between the length of a deletion and the age of the defunct element in which it is found. On the other hand, if $|4N_e s| \approx 1$ or less across the whole observed range of deletion sizes, then no such correlation is expected. Figure 4 shows that there is actually no correlation between the length of deletions and the age of the elements (Spearman's $r_s = 0.06$, $P = .62$). We have also tested whether the smallest deletions (1–10 bp) are different in average age compared with the largest deletions (77–432 bp), and here again there is no detectable effect (Wilcoxon test, $P = .63$). The same finding holds for deletions in the *D. virilis* dataset (Petrov et al. 1996; Petrov and Hartl 1997).

We conclude that genome-wide selection for a compact genome, if it occurs, is not strong enough to differentially affect the persistence times of *Drosophila* deletions ranging in length from 1 to 400 bp, suggesting that our sample of deletions in defunct retroelements affords a virtually unbiased estimate of the rate and size distribution of spontaneous deletion formation.

Summary

Studies of pseudogene evolution are important in that they may yield estimates of the rates and patterns of spontaneous mutation. The operational definition of a pseudogene is a gene duplication that has no protein-coding function, and that shows no evidence of selective constraints on its nucleotide sequence. Pseudogenes are presumed to accumulate various types of substitutions by random genetic drift, unbiased by natural selection, and thus are thought to reflect the underlying patterns of spontaneous mutation.

In this article we discuss a possibility that insertions and deletions (indels) in pseudogenes may be subject to selection pressure resulting from their incremental effects on total genome size (Charlesworth 1996; Petrov et al. 1996; Petrov and Hartl 1997). For example, if there is a genome-wide selection pressure toward a more compact genome, then one might expect that deletions, especially longer ones, would be overrepresented among observed mutations, relative to nucleotide substitutions and even more so relative to insertions.

A genome-wide selection hypothesis was prompted by our recent findings that *Drosophila* pseudogene-like sequences lose DNA through frequent deletions much faster than mammalian pseudogenes, which we interpreted as reflecting intrinsic differences in the rate and size distribution of spontaneous deletions (Petrov et al. 1996; Petrov and Hartl 1998). An alternative interpretation based on genome-wide selection for a more compact genome in *Drosophila* was also suggested (Charlesworth 1996).

In this article we argue that selection for smaller genome size is very unlikely to be the sole agent responsible for the difference in deletion spectra between *Drosophila* and mammals, and also is very unlikely to be intense enough to bias the frequency of deletions observed in *Drosophila.* Our analysis is based on theoretical considerations showing that the intensity of genome-wide selection for a more compact genome acting on any indel must be linearly proportional to its indel length, as long as the indel is a very small fraction of the total genome size.

Among pseudogenes in mammals and defunct retrotransposons in *Drosophila,* the observed number of deletions, relative to the observed number of nucleotide substitutions, is approximately equal for the size range of 1–5 bp, but much larger in *Drosophila* for size classes 6–10 bp, 11–15 bp, and 16–30 bp. Surprisingly the magnitude of the excess in each of the categories of larger deletions does not increase according to the size of the deletion, as would be expected if genome-wide selection for a compact genome were acting on deletions in proportion to their size. Furthermore, any assumption of weak selection favoring deletions as small as those actually observed necessarily implies a much larger intensity of selection acting against the persistence or fixation of newly arisen defunct retroelements in the first place, but strong deleterious effects are inconsistent with our observations of very long persistence times and even fixation of such elements in *Drosophila* species. Furthermore, the persistence times of deletions ranging from 1 to 400 bp are not significantly different, which is unexpected in a model of genome-wide selection for a compact genome.

We therefore conclude that the frequency and size distribution of small indels found in mammalian pseudogenes and defunct retroelements in *Drosophila* do not appear to be biased to any significant extent by genome-wide directional selection acting on total genome size. This does not mean to imply in any way that natural selection is not acting on genome size in *Drosophila,* but rather that such selection, if present, is not strong enough to noticeably bias patterns of small deletions/insertions (1–400 bp). This justifies the use of these sequences to estimate the underlying rates and patterns of spontaneous nucleotide substitution and indel formation. The analysis of pseudogenes and defunct retroelements promises to yield rich new information about the basic properties and patterns of spontaneous mutation, and about how these patterns may change in organisms over evolutionary time.

References

Akashi H, 1997. Codon bias evolution in *Drosophila*: population genetics of mutation-selection drift. Gene 205:269–278.

Beale D and Lehmann H, 1965. Abnormal hemoglobin and the genetic code. Nature 207:259–261.

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

Charlesworth D, Charlesworth B, and Morgan MT, 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141:1619–1632.

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

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

Li W-H, Gojobori T, and Nei M, 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239.

Li W-H, Luo C-C, and Wu C-I, 1985. Evolution of DNA sequences. In: Molecular evolutionary genetics (MacIntyre RJ, ed). New York: Plenum; 1–94.

Li W-H, Wu C-I, and Luo C-C, 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions and its evolutionary implications. J Mol Evol 21:58–71.

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

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

Petrov DA and Hartl DL, 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc Natl Acad Sci USA 96:1475–1479.

Petrov DA, Lozovskaya ER, and Hartl DL, 1996. High intrinsic rate of DNA loss in *Drosophila.* Nature 384: 346–349.

Sharp PM and Li W-H, 1989. On the rate of DNA sequence evolution in *Drosophila.* J Mol Evol 28:398–402.

Weiner AM, Deininger PL, and Efstratiadis F, 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661.

Wu CT and Morris JR, 1999. Transvection and other homology effects. Curr Opin Genet Dev 9:237–246.

Zuckerkandl E, Derancourt J, and Vogel H, 1971. Mutational trends and random process in the evolution of informational macromolecules. J Mol Biol 59:473–490.

Corresponding Editor: Masatoshi Nei