Theoretical Population Biology **61**, 533–546 (2002)
doi:10.1006/tpbi.2002.1605

TPB

Mutational Equilibrium Model of Genome Size Evolution

Dmitri A. Petrov

Department of Biological Sciences, Stanford University, 371 Serra Street., Stanford, California 94025
E-mail: dpetrov@stanford.edu

Received February 25, 2002

The paper describes a mutational equilibrium model of genome size evolution. This model is different from both adaptive and junk DNA models of genome size evolution in that it does not assume that genome size is maintained either by positive or stabilizing selection for the optimum genome size (as in adaptive theories) or by purifying selection against too much junk DNA (as in junk DNA theories). Instead the genome size is suggested to evolve until the loss of DNA through more frequent small deletions is equal to the rate of DNA gain through more frequent long insertions. The empirical basis for this theory is the finding of a strong correlation and of a clear power-function relationship between the rate of mutational DNA loss (per bp) through small deletions and genome size in animals. Genome size scales as a negative 1.3 power function of the deletion rate per nucleotide. Such a relationship is not predicted by either adaptive or junk DNA theories. However, if genome size is maintained at equilibrium by the balance of mutational forces, this empirilical relationship can be readily accommodated. Within this framework, this finding would imply that the rate of DNA gain through large insertions scales up a quarter-power function of genome size. On this view, as genome size grows, the rate of growth through large insertions is increasing as a quarter power function of genome size and the rate of DNA loss through small deletions increases linearly, until eventually, at the stable equilibrium genome size value, rates of growth and loss equal each other. The current data also suggest that the long-term variation is genome size in animals is brought about to a significant extent by changes in the intrinsic rates of DNA loss through small deletions. Both the origin of mutational biases and the adaptive consequences of such a mode of evolution of genome size are discussed.

Key Words: **genome size; mutational bias, deletions; insertions, junk DNA.**

INTRODUCTION

It is tempting to think of genomes as aggregates of genes. Although this description is quite fair for prokaryotes and viruses, it is woefully inappropriate for eukaryotes. In most eukaryotes, genes are in a clear minority, with most of the DNA devoted to nongenic, largely unconstrained, often repetitive DNA. Moreover, eukaryotes vary greatly in the amount of noncoding DNA, producing a remarkable, 200,000-fold variation in genome size (Gregory and Hebert, 1999).

Evolution of genome size begins with mutations affecting the length of noncoding DNA. These mutations change genome size to the extent that they can get past the vagaries of random genetic drift and effects of natural selection and reach fixation in populations. In addition, enzymes and genetic elements producing length mutations also evolve by drift and natural selection. Given the variety of different types of nongenic sequences and of the forces possibly affecting their abundance and length, it is understandable that evolution of genome size remains enigmatic (Gregory, 2001; Petrov, 2001).

One way to approach this multifaceted problem is to try to measure the forces affecting genome size. In this paper, I will discuss the measurements of one such force—the rate and pattern of small deletions and

0040-5809/02 $35.00

insertions in different organisms. The key finding is that, surprisingly, this one parameter statistically explains a significant amount of the variation in genome size in the studied organisms. I will discuss the implications of this result and will propose a mutational equilibrium model of genome size evolution.

PATTERN, EVOLUTION, AND CONSEQUENCES OF DELETION/ INSERTION BIASES

The first obvious step in investigating insertion/deletion (indel) rates is their measurement. However, because mutations in general and indels in particular are very rare, they are often difficult to measure with precision in a laboratory setting. A common alternative approach is to study substitutions in pseudogenes (Li *et al.*, 1981; Gojobori *et al.*, 1982). Pseudogenes are nonfunctional copies of genes, devoid of coding function. Because mutations in pseudogenes are not selected based on their effect on the coding capacity of the sequence, they are often assumed to be truly neutral. This assumption allows the pattern of substitutions in pseudogenes to be used as a proxy for the mutational spectrum (relative rates of different types of mutations).

The usefulness of pseudogenes is limited by two problems. First, pseudogenes are rare in some organisms, such as Drosophila. In others, less well-studied organisms, pseudogenes may be present in large numbers, but require much preliminary work for their identification. As a result, large numbers of pseudogenes have been available in only a few pseudogene-rich, well-studied genomes (mostly in mammals, now also in *Caenorhabditis elegans* (Robertson, 1998; Robertson, 2000; Harrison *et al.*, 2001)). In addition, pseudogenes are not truly inert—they are likely to exert phenotypic effects simply through their bulk, effects on nearby genes, and sometimes expression of partial, possibly deleterious products. Mutations that interfere with these effects may be subject to selection (Robin *et al.*, 2000). Moreover, the strength of selection may vary for different types of mutation, biasing our sample in an unknown way.

A partial solution to both of these problems can come from studying other, pseudogene-like sequences, such as dead copies of transposable elements (in particular, nonLTR retrotransposable elements) or nuclear insertions of mitochondrial DNA (numts). These sequences are abundant in most organisms (Malik *et al.*, 1999; Bensasson *et al.*, 2001) and are easy to clone. Therefore, they can provide us with a ready source of pseudogene-like sequences even in poorly studied organisms. Moreover, by combining a large number of different kinds of pseudogenes, with different sequences, genomic locations, copy number, and ages, we can hope to average out peculiar selective effects in particular pseudogenes (Petrov, 2002). We can further test for such effects by assessing heterogeneity of the inferred mutational biases. Note, however, that selection acting on mutations at a genomic level should affect all pseudogenes to a similar extent and would need to be assessed separately (Petrov and Hartl, 2000; Petrov, 2002).

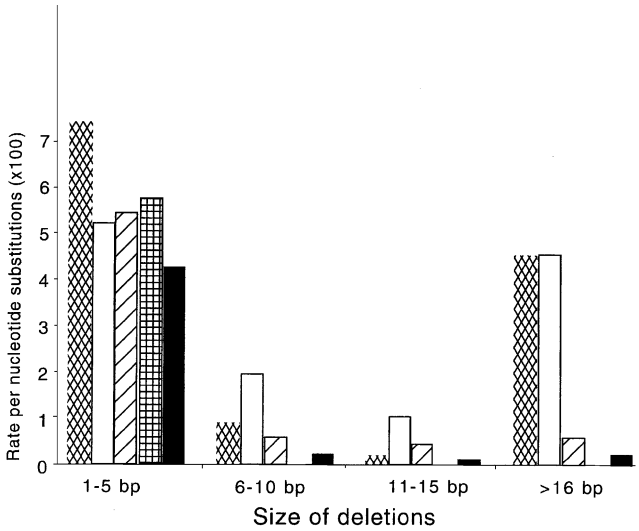

FIG. 1. Relative rates of deletions of different sizes and nucleotide substitutions in *C. elegans* (angled crosshatched bars), *Drosophila* (open bars), *Laupala* (straight crosshatched bars), mammals (black bars), *Podisma* and *Italopodisma* grasshoppers (crisscross hatched bars).

In the rest of the paper, I will assume that differences in the indel spectra in pseudogenes are generated through global, genome-wide differences in the mutational processes and not through differential effects of selection in different organisms (Charlesworth, 1996). The full justification for this assumption can be found elsewhere (Petrov and Hartl, 2000; Petrov, 2002).

CURRENT ESTIMATES OF THE SHORT INDEL SPECTRA

A variety of pseudogene-like sequences (including *bona fide* pseudogenes, dead copies of nonLTR ele-

TABLE I

Rate of DNA Loss through Small Indels and Genome Size

	Drosophila[a]	*C. elegans*[b]	*Laupala* crickets[c]	Mammals (primates and rodents)[d]	*Podisma* grasshoppers[e]
Genome size (Mbp)	179	100	1910	~ 3000	18150
Size of the data sets[f]	669p, 87d, 10i	68d, 16i	662p, 45d, 14i	2662p, 132d, 27i	169p, 12d, 6i
Average rate of deletions per insertion	8.7	4.0	3.2	5.0	2.0
Average rate of deletions per bp substitution	0.13	n.d.[g]	0.07	0.05	0.06
Average rate of insertions per bp substitution	0.015	n.d.	0.02	0.01	0.03
Average size of deletions	35	48	7.0	3.2	1.6
Average size of insertions	2.9	10	6.5	2.4	1.2
Average rate of DNA loss (bp/per 1 bp substitution)[i]	4.5	5.9[h]	0.34	0.13	0.06

[a] Data from Petrov and Hartl (1998).
[b] Data from Robertson (2000).
[c] Data from Petrov *et al.* (2000).
[d] Data from Graur *et al.* (1989).
[e] Data from Bensasson *et al.* (2001).
[f] p = nucleotide substitutions; d = deletions; i = insertions.
[g] Not determined.
[h] Assuming the rate of indel/point substitution as observed in Drosophila. Differences between the rates of DNA loss in Drosophila and *C. elegans* are not very robust because of the flat distribution of deletions of larger than 15 bp in both organisms.
[i] This is measured as $D = (\text{rate of deletions per nucleotide substitution}) * (\text{average size of deletions}) - (\text{rate of insertions per nucleotide substitution}) * (\text{average size of insertions})$.

ments, and numts) have now been used to estimate rates of small (less than 400 bp) deletion and insertions in *Drosophila* flies, *Laupala* crickets, *Podisma* & *Italopodisma* grasshoppers, mammals and *C. elegans* (Graur *et al.*, 1989; Gu and Li, 1995; Petrov *et al.*, 1996, 1998, 2000; Ophir and Graur, 1997; Robertson and Martos, 1997; Petrov and Hartl, 1998, 2000; Lozovskaya *et al.*, 1999; Robertson, 2000; Bensasson *et al.*, 2001; Petrov, 2002). These results are summarized in Fig. 1 and Table 1.

From this admittedly limited research, several tentative patterns emerge. In all cases indels are rare relative to point substitutions (nucleotide substitutions are from 7 to 17 times more numerous than indels). Also across the board, deletions are more frequent and larger than insertions. Thus among indels smaller than 400 bp there is a mutational pressure toward DNA loss. The magnitude of this pressure varies dramatically—by almost 100-fold—primarily because of variation in the rate of deletions larger than 5 bp (5–400 bp). Indeed, whereas insertions are invariably short and rare, and the smallest deletions (1–5 bp) have similar rates in all cases (Fig. 1) (Graur *et al.*, 1989; Ophir and Graur, 1997; Robertson, 2000; Bensasson *et al.*, 2001), the larger deletions (> 5 bp and especially > 15 bp) are common in some organisms (*Drosophila* and *C. elegans*) and virtually absent in others (*Podisma* grasshoppers).

EVOLUTION OF THE SHORT INDEL SPECTRA

The observed pattern suggests that the only variable parameter of the short indel spectra is the size distribution of deletions and, in particular, presence or absence of deletions longer than 5 bp. What could be the reasons, if any, for this pattern?

Let us first consider the mechanisms responsible for indel formation. Many small indels are thought to be a product of replication slippage, schematically shown in Fig. 2 (Albertini *et al.*, 1982; Kunkel, 1986, 1990; Bebenek and Kunkel, 1990; Ling *et al.*, 2001). Deletions are generated when the replication complex skips across a number of nucleotides and fails to replicate them, whereas insertions are formed when the same region is mistakenly re-replicated. This mechanism generates a thermodynamic asymmetry. Long insertions require melting of a long stretch of already replicated DNA, whereas deletions do not. This could be one mechanistic reason for the rarity of long insertions generated through replication slippage.

There is no obvious mechanistic reason why the rate of occurrence of indels should be relatively constant across organisms. However, a consideration of deleterious effects of deletions and insertions on genes could

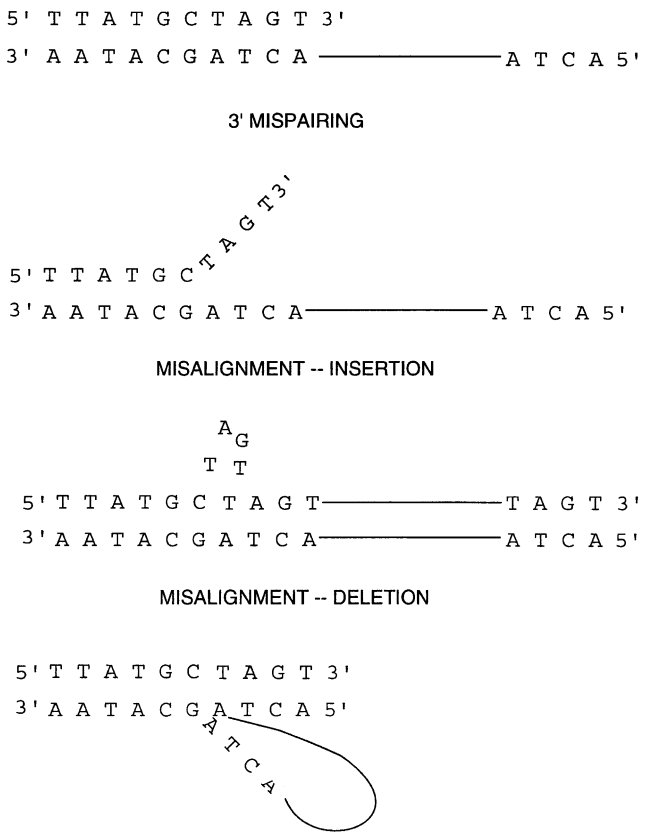

FIG. 2. Both deletions and insertions can be generated through slippage of DNA polymerase during replication. Both events start by unpairing of a stretch of nucleotides at the 3′ end of the newly synthesized strand. Rereplication of this stretch would result in an insertion of a number of nucleotides into the replicated strand. The 3′ unpairing could also lead to incorrect pairing further downstream on the template strand, resulting in a deletion. In this model, insertion length is limited by the extent of initial unpairing, whereas deletion length is limited by the accessibility of the downstream template strand. Differences in the distance between the 3′ end of the newly synthesized strand and the replication fork in different organisms may lead to differences in the average size of deletions.

provide a possible reason for their invariably low rates. Unlike nucleotide substitutions, indels are almost invariably deleterious when they occur within coding regions. It is thus possible that selection against a high rate of deleterious mutation maintains a relatively low rate of indels relative to nucleotide substitutions.

But why then would the size of deletions remain so variable? Why would not deletion size be similarly affected by selection against high deleterious mutation rate? The answer may be that although indels in coding regions are almost invariably disruptive, this effect is at best only weakly dependent on size. Indeed, since both 2 and 20 bp deletions in a coding region are likely to produce protein-inactivating frameshifts, both kinds should be similarly deleterious to genes. Thus, it is possible that variation in the average size of deletions may not be noticed by selection for lower deleterious mutation rate. At the same time, changes in the size of deletions can profoundly affect the average rate of loss of nongenic, unconstrained or very mildly constrained DNA.

As an example consider the indel spectra in *Laupala* crickets and *Podisma* grasshoppers (Table I). Due to the larger average deletion size, the rate of nonfunctional DNA loss per point substitution is ~ 6 times higher in *Laupala*. However, the rate of indels is virtually identical in both organisms, leading to an unchanged rate of deleterious mutation rate due to indels. Compare this with a scenario where *Laupala* acquires the same 6-fold faster DNA loss through the increase of the rate of deletions but not their size. In this case its deleterious mutation rate would go up by ~ 3-fold.

Putting all of these considerations together, it seems that natural selection for lower deleterious

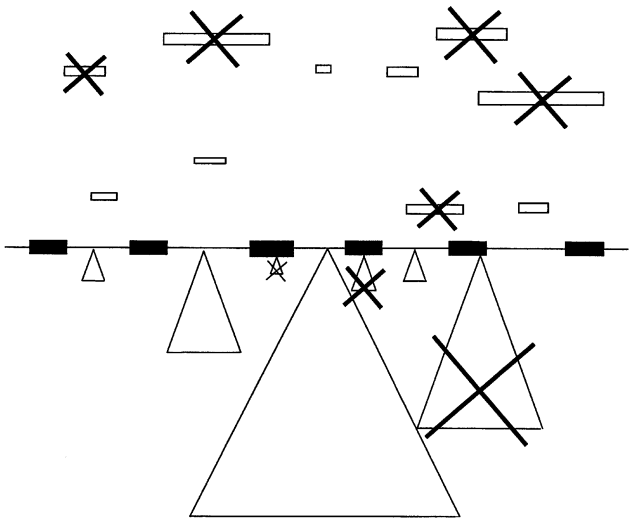

FIG. 3. Genes are depicted as black boxes and intergenic sequences as a line connecting them. Insertions (triangles) inactivate genes with the same probability independently of their length, whereas longer deletions (open bars) tend to inactivate genes more often.

mutation rate keeps deletions and insertions rare, while thermodynamics of DNA replication may keep insertions but not deletions very small. The only parameter that is free to vary is the size of small deletions. This also means that at a small scale (1–400 bp) mutation is unlikely to be biased toward DNA gain in most organisms.

LONG DELETIONS AND INSERTIONS

So far we have been considering indels that are smaller than average genes (1–400 bp). For such small indels the parameter relevant to the deleterious mutation rate is the rate of occurrence but not their size. This is not true for larger indels, whose size is of the same magnitude or larger than genes, because here we also have to consider the possibility of edge effects.

Consider a genome that is a mixture of genic, functional DNA (genes and nearby regulatory regions) separated by stretches of noncoding DNA (Fig. 3). Because deletions have two breakpoints, large deletions would affect genes more frequently than short deletions. At the extreme, deletions longer than all intergenic spacers are invariably deleterious. On the other hand, insertions, having only one breakpoint, would have the same probability of gene inactivation independently of their length.

There are two consequences of this asymmetry between deletions and insertions. The amount of noncoding DNA, and therefore genome size, is unlikely to be affected significantly by long deletions. These are almost always deleterious and will be removed from the genome by natural selection. This is particularly true for very compact genomes with short intergenic regions; as genomes and noncoding regions grow in size, large deletions will become relatively more important. The importance of large deletions may also be higher in the genomes with an uneven gene distribution—long regions devoid of genes could accept long deletions. Large insertions, on the other hand, may have a very strong impact on genome size for both compact and large genomes. The abundance of long insertions of transposable elements, phage and organel DNA in eukaryotic genomes all speak to the power of long insertions.

In addition this means that selection for lower deleterious mutation rate should reduce the rate of long deletions (especially in compact genomes), but not necessarily that of long insertions. Unfortunately, at this time, we do not have adequate information to test this prediction.

EVOLUTION OF GENOME SIZE

A 100-fold difference in the strength of the small indel bias toward DNA loss raises the possibility that it might be important in genome size evolution. How conceivable is this? One argument against this possibility is that the rate of DNA loss is extremely slow, even in *Drosophila* or *C. elegans*. How could such a weak force be of any relevance?

This argument is entirely correct in cases where we need to explain fast changes in genome size. For instance, the growth of the maize genome by 2-fold in the past 3 Myr (SanMiguel *et al.*, 1996, 1998) is unlikely to be caused by changes in the strength of the indel bias. A much faster force must have been in play here, in this case most likely a sharp increase in TE activity (SanMiguel *et al.*, 1996, 1998). In other cases of rapid changes in genome size, we might expect to find natural selection or a fast mutational force (TE mobilization, expansion of simple repeats, polyploidization, etc.) as the culprit. However, when we consider the long-term evolution of genome size, over hundreds of Myr, slow and persistent indel bias may be just as efficacious as any fast but sporadic force (Petrov, 1997, 2001).

There is another reason to take indel biases seriously. The long-term pattern of genome size evolution

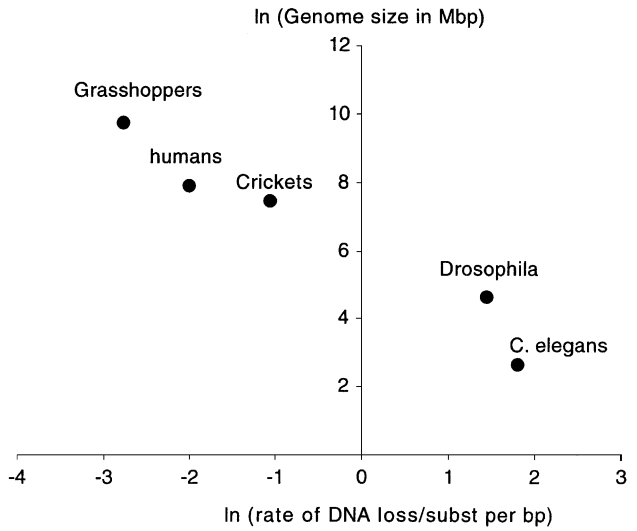

FIG. 4. Negative correlation (depicted on a ln/ln scale) between the rate of DNA loss through small deletions (per nucleotide per nucleotide substitution) and the genome size.

demonstrates that genomes shrink or expand at all levels of their organization. Larger genomes tend to have longer introns (Vinogradov, 1999) and inter-enhancer regions (Bergman and Kreitman, 2001), more pseudogenes and simple repetitive DNA. This pattern implies the action of a global force capable of affecting all sequences in the genome. Along with natural selection, indel bias is a prime candidate for such a global force (Petrov, 2001).

NEGATIVE CORRELATION BETWEEN INDEL BIASES AND GENOME SIZE—PATTERN AND CAUSATION

If indel biases are important in genome size evolution, we should see a correlation between indel biases toward DNA loss and the amount of nongenic DNA (e.g., excluding coding and regulatory genic sequences). In large genomes such as found in grasshoppers or humans, genes are in such a minority that the size of the nongenic part is probably close to the overall genome size. In the compact *D. melanogaster* or *C. elegans* genomes this is clearly not the case, however. In *D. melanogaster* we can estimate that $\sim 35\%$ of euchromatin and $\sim 85\%$ of heterochromatin are nongenic and unconstrained, giving us ~ 90 Mbp of nongenic DNA and 86 Mbp of genic DNA (Shabalina and Kondrashov, 1999; Bergman and Kreitman, 2001; Shabalina *et al.*, 2001). For simplicity, we can assume that all five organisms have the same amount of functionally constrained, genic DNA. We can then obtain the size of the unconstrained genomes by subtracting 86 Mbp from all values of genome size.

The resulting relationship is shown in Fig. 4. There is a strong negative correlation between the size of the nongenic part of the genome and the strength of indel biases toward DNA loss. The non-parametric Kendall's test of rank correlation is significant ($P = 0.05$). There also appears to be a power–function relationship between genome size and the rate of DNA loss. Pearson correlation coefficient with ln-transformed data is $r = -0.98$, which is significant at $P = 0.004$. The mathematical relationship has the following power–function form

$$G \propto D^{-1.3}. \quad (1)$$

In this formula, G is the nongenic or unconstrained portion of genome size (Mbp) and D is the mutational rate of DNA loss through small indels per nucleotide per nucleotide substitution (i.e., $D =$ (rate of deletions per nucleotide substitution)*(average size of deletions) −(rate of insertions per nucleotide substitution)*(average size of insertions)).

The data are admittedly very limited, but very suggestive. Assuming that this pattern holds in the future, what lessons can we draw from it? In particular, can we infer causation from this negative correlation?

I believe yes, but tentatively so. *Prima facie*, other things being equal, changes in indel biases can change genome size over time. The faster the DNA loss, the smaller the genome size should be. Thus there is a plausible causal mechanism that could underlie the observed negative correlation. The reverse causation, from genome size changes through other mechanisms (natural selection, activity of TEs, polyploidization, etc.) to indel biases appears to be less plausible. At least at the moment we have no evidence that genome size per se affects indel biases. Measurements of indel biases in closely related organisms with drastically different genome sizes, where genome size change cannot have been brought about by shifting indel biases, could shed light on this issue. If genome size directly affects indel biases, we should see negative correlation between indel biases and genome size in such cases as well.

There is also a possibility that this pattern is coincidental or unrelated to genome size evolution. This is a very general worry, applicable to practically any empirical observation. Although we cannot rule coincidence out, we will concentrate on direct causal hypotheses.

ARE GENOME SIZES IN EQUILIBRIUM?

In the rest of the discussion I will only consider scenarios where genome size is at or close to a long-term equilibrium. In general this assumption is clearly not valid—it is sufficient only to think of evolution of genome size in maize. It is hard to think of the maize genome as being in an equilibrium, given that it grew by 100% in 3 Myr (SanMiguel *et al*., 1996).

However, in our discussion two considerations make this more reasonable. First, over very long periods of time the averaged effect of multiple forces, including slow and steady ones, such as mutational DNA loss, and fast and sporadic ones, such as bouts of transposition, may drive the genome size close to a long-term equilibrium value. In this picture, recent presence or absence of bouts of transposition such as observed in maize would appear as noise around the long-term equilibrium value. (For instance, the fact that all grasses have genome sizes larger than that in *Arabidopsis*, may be due to the long-term forces; difference in genome size between maize and *Sorghum* would not be and will appear as noise.) Second, the fact that we observed such a clear mathematical relationship between the rate of DNA loss and genome size is *prima facie* evidence that genome sizes are close to their long-term equilibria in the studied organisms. We can alternatively say that the differences among the long-term equilibrium genome sizes swamp out the noise of recent, and quick changes in genome size at least in these organisms.

CAN THE CURRENT MODELS EXPLAIN THE OBSERVATION?

In the rest of the paper we will discuss the following evolutionary scenario. At some point in time, through drift or natural selection, proteins involved in DNA replication, recombination, or repair change and start generating a new pattern of indels that is more or less biased toward DNA loss. The questions that we will consider are (i) what should happen in such a case according to the current theories of genome evolution, and (ii) what are the conditions under which these changes eventually would lead to the establishment of the observed negative, power–function relationship between indel biases and genome size (Fig. 4 and Formula (1)). In other words, we imagine that indel biases changed in the ancestors of the studied organisms and produced the observed 100-fold variation in genome size. If this is what happened, how can we make sense of it?

CURRENT MODELS

Roughly speaking, currently there are two types of theories of genome size evolution—adaptive and junk DNA theories. Junk DNA theories maintain that larger genome sizes are generally maladaptive because they add energetic costs, lengthen replication time, and have other deleterious phenotypic effects. Despite the selective costs, large genomes evolve because of the persistent pressure of junk DNA addition as a byproduct of pseudogene formation and selfish DNA (transposable element) activity (Ohno, 1972; Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Adaptive theories, on the other hand, posit that large genome sizes are often adaptive through their effect on phenotypic characters such as cell size, nucleus size, rates of basal metabolism, seed size, cell division rates and many others (Gregory and Hebert, 1999). Genome size then evolves under positive selection pressure to fit the physiological and ecological needs of different organisms and is maintained at equilibrium by stabilizing selection.

ADAPTIVE THEORIES

Imagine that adaptive scenarios are correct in that genome size evolves to fit the physiological and ecological needs of the organism. On this model, the phenotypic needs of the organism determine genome size. In equilibrium, both decreases and increases in genome size are deleterious and the genome size is maintained by stabilizing selection. If stabilizing selection is very strong, then our imagined change in indel biases should not markedly affect genome size. It might shift the population average somewhat, as a result of the mutation–selection balance, but should not result in a directional long-term change in genome size. It certainly would not be expected to yield a 100-fold change in genome size.

This does not mean of course, that selection does not act on genome size or that genome size does not exert adaptively important influence on the phenotype. It is only that the apparent responsiveness of genome size to changes in indel biases would make us doubt that the stabilizing selection acting on genome size is very strong.

It might mean, for example, that the reverse adaptation—the rest of the organism adapting to its genome size through co-evolution of other characters and habitat change—is a likelier model.

JUNK DNA THEORIES

On the junk DNA view, larger genome sizes are maladaptive as a rule and attained because of persistent "junk" DNA addition. The original paper by Ohno (1972) did not make clear what would prevent unchecked expansion of the junk DNA. Later papers, which also identified transposable elements as the main engine of "junk" DNA addition, suggested that purifying selection against maladaptive genome growth could stop the growth (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). This is the version of "junk" DNA theory we will consider here.

In order for purifying selection to stop genome expansion, the fitness function of genome size has to be concave, such that persistent DNA addition increases genome size while fitness function is flat and stops when it drops sharply. At equilibrium, selection against additional DNA is strong enough to counterbalance mutational pressure toward DNA gain. Effectively, although activity of transposable elements and creation of pseudogenes is still much faster than creation of deletions, insertions reach fixation at a much lower probability than deletions. At equilibrium, genome size does not change because the product of the low mutation rate of deletions multiplied by their high rate of fixation is equal to the product of the high mutational rate of insertions and their low rate of fixation.

What would happen if we change, say decrease, the mutational rate of deletions in this system? It all depends on how sharply the fitness function drops off. If it drops off very sharply we expect to see no effect. In this case even a slight increase of genome size would drastically increase the strength of selection against additional genome size growth and would maintain genome size at a virtually unchanged level. If fitness function does not drop off precipitously, then genome size will start growing until the new equilibrium is reached.

Let us now consider conditions that would produce the observed relationship between the mutational bias toward DNA loss and genome size (Fig. 4 and Formula (1)). It is clear that the strength of selection has to change in a very particular way in order for this to happen. One can see (the derivation is in the appendix) that the selection coefficient associated with a 1 bp indel α, has to have the following relationship with the effective populations size (N_e), the rates (i and d) and average lengths (L_d and L_i) of deletions and insertions, the genome size (G), and the coefficient of proportionality (k) between genome size and rate of DNA loss per bp in (1):

$$\alpha \approx 1/(4N_eL_i)(\ln(iL_i^2) - \ln(kG^{-1.3}L_d) \qquad (2)$$

in order to generate the observed negative correlation. This is not impossible, but we find it unlikely. Given the wide variation of selective needs and population sizes of the studied organisms, there seems no reason to suspect that this relationship would be faithfully maintained in evolution. Selection coefficients should depend on ecological and physiological needs of the particular organism. It is unclear why they should also depend on the other parameters in (2) and why this functional relationship should be maintained across organisms with very different physiological and ecological needs. On balance, versions of junk DNA theory in which natural selection against genome growth counterbalances the intrinsic tendency of genomes to expand have difficulty naturally explaining the observed empirical relationship (Fig. 4 and Formula (1)).

MUTATIONAL HYPOTHESIS OF GENOME SIZE EVOLUTION

If it is not selection that determines the equilibrium genome size, what else could it be? As we already discussed, the mutational pressure at the level of small indels is always biased toward DNA loss. If the preferential fixation of small deletions over small insertions is not prevented by selection, it means that all genomes are constantly losing DNA through small indels. Because genomes are not at their smallest values across the board (even in a compact *Drosophila* or *C. elegans* genomes), it means that DNA loss through small indels has to be counterbalanced by DNA gain through the preponderance of large insertions over large deletions.

Let us consider this scenario. Ignoring selection, the loss of DNA at a small scale is then simply the product of the rate of DNA loss per nucleotide (D) and the number of nucleotides in the unconstrained portion of

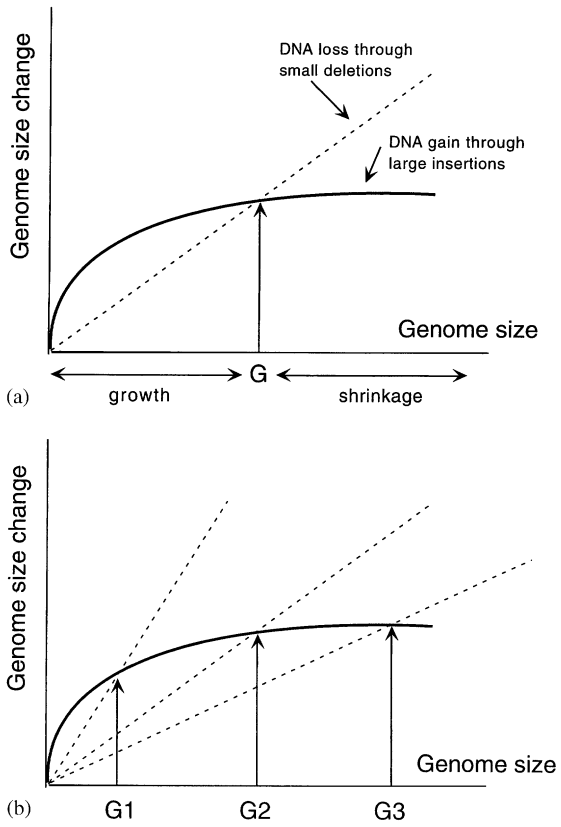

FIG. 5. Establishing mutational equilibrium genome size. Rate of DNA loss through small deletions scales linearly with genome size (dashed lines), whereas DNA gain through large insertions scales slower than linear (solid line). (a) The equilibrium genome size is established when the rate of DNA gain is equal to that of loss (the intersection of the two lines). (b) Different rates of DNA loss per nucleotide lead to different equilibrium genomes sizes ($G1, G2, G3$).

genome size (G):

$$\text{Rate of DNA loss per genome} = DG. \quad (3)$$

At equilibrium, the rate of DNA gain (I) through the imbalance of large insertions over large deletions has to be equal to the rate of DNA loss through the imbalance of small deletions over small insertions:

$$DG = I. \quad (4)$$

If we now use the empirical relationship between the rate of DNA loss per bp and genome size (1), we infer the following relationship for I:

$$I \propto G^{1/4}. \quad (5)$$

Eq. (5) means that if the rate of DNA gain through large insertions increases roughly as a quarter-power function of genome size, because the rate of genome size loss through small deletions increases linearly with genome size, over long time we will obtain the observed negative correlation between the genome size and the rate of small indels (Fig. 4 and Formula (1)).

This relationship is shown in Fig. 5a. For small genome sizes the rate of genome size increase is higher than that of DNA loss resulting in genome size growth. However, since the rate of DNA loss through small deletions grows linearly and thus faster than the rate of DNA gain, for very large genome sizes DNA loss is faster than DNA growth. Therefore, there exists a stable equilibrium at a finite value of genome size. Importantly, any concave increase in the rate of DNA gain as a function of genome size

$$I = aG^b, \quad \text{where } b < 1 \quad (6)$$

would generate a stable genome size equilibrium with $G \geqslant 0$. This is the essence of the mutational equilibrium hypothesis for the evolution of genome size.

In the five studied organisms, much of the evolution of different equilibrium genome sizes can be attributed to the rates of DNA loss per bp through the preponderance of small deletions over small insertions (Fig. 5b). In other cases, however, the relationship (6) can change as well (change in the parameters a and b). In the five studied organisms, parameters a and b apparently have been less variable compared to the rate of DNA loss per bp through small indels allowing us to see relationship (1). However, even in these organisms DNA loss through small indels does not explain all of the data (G-test, $p \ll 0.01$). The remaining variation is probably due to the variation in the processes of large insertion (i.e., parameters a and b) and/or due to the recent nonequilibrium changes of genome size (similar to that observed in maize (SanMiguel *et al.*, 1996, 1998).

DNA GAIN THROUGH LARGE INSERTIONS AND GENOME SIZE

Unlike the junk DNA and adaptive theories of genome size, mutational equilibrium does not depend on natural selection acting at the level of genome size. It requires instead that the genomic rate of DNA addition should increase with genome size at a slower than linear rate. Why should that be so?

Large insertions are comprised of both selfish, transposable elements (TEs) and nonselfish DNA insertions such as tandem duplications, pseudogenes, organel DNA insertions, and many others. First let us consider TEs. The activity of transposable elements (TEs) is probably the primary source of large insertions in eukaryotic genomes. As genomes grow the number of transposable elements grows as well (Kidwell, 2002), and, it would be natural to assume, so does the time-average genomic rate of transposition. The number of fixations of TEs should also grow with genome size as there will be more and more nondeleterious insertion sites. We should expect then that the rate of TE DNA addition should be a monotonically increasing function of genome size.

The increase of transposition rate in the long term should also be slower than linear as a function of genome size. Indeed, if it is faster than linear, then the copy number of TEs would increase exponentially and genome sizes would grow infinitely and uncontrollably. Many empirical and theoretical studies have considered this question (Nuzhdin, 1999). In some cases, the reduction of transposition rate is a function of a reduced activity of the transposase as its concentration rises (Hartl *et al*., 1997). In other cases, such as in hybrid dysgenesis (Kidwell *et al*., 1977; Petrov *et al*., 1995), the increased activity of TEs leads to the establishment of genomic immunity and the reduction of transposition rates in a few generations. The probability of fixation of individual copies of TEs can also go down with the increased rate of transposition, if TEs are synergistically deleterious in their effect (Charlesworth and Langley, 1989). This could happen, for example, if ectopic recombination among dispersed copies of TEs leads to deleterious effects or if TE-produced proteins are disruptive for the cell in a nonlinear fashion. The increased rate of transposition would also lead to a likelier evolution of cellular modifiers of transposition and their presence at a higher frequency in populations (Nuzhdin, 1999).

Consideration of large deletions also suggests that the rate of DNA addition at a large scale should diminish in larger genomes. Because average distances among genes increase as genome size grows, large deletions become less deleterious on average in larger genomes and thus should have a greater impact on genome size evolution.

The rate of nonselfish DNA addition should also be an increasing, linear or concave function of genome size. It should increase along with the number of nondeleterious insertion sites (important for processed pseudogene formation and insertions of organel DNA). The decrease in the density of genes should make larger tandem duplications less deleterious as they will contain fewer genes on average and will have a smaller effect on gene dosage, but will also make neutral large deletions likelier.

On balance, our understanding of forces affecting large-scale DNA addition and removal is consistent with the genomic rate of DNA addition increasing with genome size monotonically and at a slower than linear rate. We do not, however, have specific expectations about the exact shape of this relationship (i.e., the inferred quarter-power function) or understand why this relationship should remain approximately constant across the studied organisms (from mammals to insects to worms). One possibility is that the constancy of this relationship is related to the overall similarity of types of TEs in these animals. However, we predict that wider taxon sampling (for instance in plants) will demonstrate the existence of different relationships between genome size and rates of DNA addition through large insertions.

ADAPTIVE ROLE OF NONGENIC DNA AND THE MUTATIONAL EQUILIBRIUM HYPOTHESIS

The current theories of genome size evolution are usually distinguished based on their view of the functional value of nongenic DNA. Whereas, adaptive theories postulate a functional role for this DNA, "junk" DNA theories propose that much of the nongenic DNA is a nonfunctional, maladaptive byproduct of pseudogene formation and transposable element activity.

Under the mutational equilibrium model, we do not need to take a stand on this issue. The claim is that even though the amount of nongenic DNA is determined through the balance of mutational forces, at any given time, this DNA may or may not have a function and may or may not be adaptive. For instance, imagine a situation where a small-genome organism evolves a much slower rate of DNA loss corresponding to a much larger equilibrium genome size. According to our model its genome will begin slowly increasing in size, with the accrued DNA likely being nonfunctional at least at the beginning. However, many genomic functions may coevolve with genome size, possibly making the new nongenic DNA essential. (Functions can "colonize" new DNA (Zuckerkandl, 2002); for instance, a larger centromeric region may result in the evolution of a

centromere definition system that depends on this large DNA amount.)

The larger genome would also affect many phenotypic characters, such as generating larger cell and nucleus sizes, slower rate of development, and so on (Gregory and Hebert, 1999). If the change is slow, we would expect the organism to adapt to this by changing its life history and modifying its physiology. On this view, genome size is of great adaptive importance. It is a nonselectively determined constraint affecting the future course of adaptive evolution.

NEUTRALITY OF GENOME SIZE

In the derivation of the mutational equilibrium hypothesis, we made the assumption that genome size is a neutral character. This assumption is clearly wrong in this extreme form. Nevertheless, relaxing this assumption may not necessarily lead to significant changes in our conclusions.

In the previous example, increase of the genome size is likely to be deleterious. However, as long as the fitness decline is small over the range of genome size variation that is produced by the segregating deletions and insertions on the way to fixation, natural selection may not be able to stem the genome size growth. If mutation rate is low and the size of individual deletions and insertions are small relative to the genome size, at any given time genome size variation in populations would be too small to produce significant selective differences. Slow and steady accumulation of DNA may result in slow, steady, and virtually imperceptible decline in fitness.

It is of course possible that in some cases there will be a sufficiently rapid decline of fitness function over small genome-size changes that would stop the genome size increase and the consequent erosion of fitness. Such rapid fitness declines may not be common, however. In many organisms, 2-fold increases in genome size over brief intervals are tolerated (SanMiguel *et al.*, 1996).

The steady decline of fitness is not the only possible outcome, however. The expected co-evolution of life-history characters should lead to the maintenance of fitness in spite of the initially deleterious genome size increase. For example, a very large genome may be disadvantageous in an annual plant. But, if along with the slow genome size increase, the plant lineage becomes perennial, the large genome may remain tolerable and in some cases even advantageous.

EVOLUTION OF THE INDEL BIAS

The potentially important effect of the small indel spectrum invites the question of the forces that would affect the indel spectrum itself. The rate and the size of small deletions and insertions depend on the function of proteins involved in DNA replication, recombination, chromatin packaging, and repair. Which forces could affect these proteins? In particular could selection at the level of genome size act through changes in the indel spectrum?

Changes in the proteins affecting indel bias are likely to have many pleiotropic effects. Other than affecting mutation rate and pattern, including the deletion/insertion bias, they could also influence such characters as rate and cost of DNA replication, DNA function in gene activity, recombination rate and many others. Overall, mutations in these proteins have a chance to fix in the population only if the overall fitness impact is either positive or at least not strongly negative.

Imagine a scenario where it is beneficial to increase the rate of replication. Changes in many of the proteins involved in DNA replication could do that. Imagine also that many such mutations have a pleiotropic effect of affecting the rate and the size of deletions and insertions at the same time. If the consequent increase in the deleterious mutation rate is too severe, such changes may not become fixed. However, we already discussed that changing the average size of deletions, without increasing their rate, would have only a minimal effect on the deleterious mutation rate. Thus of all mutations leading to the faster DNA replication, we are likeliest to see those that either do not affect deletion/insertion spectrum, or possibly only those that affect the size of deletions.

This example makes it clear that both drift and natural selection acting on proteins involved in DNA metabolism can affect deletion/insertion spectra and consequently lead to an eventual change in genome size. It is much harder, however, to imagine a scenario where a change in deletion/insertion spectrum is selected because of its effect on genome size. Such selection would depend on the linkage of the modifier of the indel spectrum and the indels themselves. As indels would occur anywhere in the genome, recombination in sexual organisms would quickly break any association. In addition, despite its long-term power, changes in indel spectra affect genome size very slowly. A mutation leading to a change in the indel spectra will not be consistently linked in populations with a sufficiently changed genome size and thus will not selected.

Consider, for example that even if *Podisma* grasshoppers evolved the *Drosophila* rate of DNA loss, leading to an eventual shrinkage of the genome size by a 100-fold, the change would result in the loss of $\sim$ 50 bp per generation. This constitutes approximately 3×10^{-9} of the *Podisma* genome. On balance, it seems extremely unlikely that selection at the level of genome size could act through changes in the indel spectra.

CONCLUSION

The extreme and counterintuitive variation in genome size among eukaryotes remains one of the key and longest-standing unsolved problems in genome biology. Attempts to resolve this problem traditionally focused on the potential function of nongenic DNA in many genomes. Whereas, some theories suggest that nongenic DNA is mostly junk accumulated maladaptively until finally checked by selection, others emphasize the potential adaptive benefits of having large amounts of DNA in certain cases. On the latter view, positive natural selection modifies genome size until it is aligned with the ecological and physiological needs of the organism.

In this paper, I propose a somewhat orthogonal view. I suggest that the equilibrium genome size may not be determined by selection acting at the level of genome size. Instead it is determined by the balance between the DNA loss through the preponderance of small deletions over small insertions, and the DNA gain through the preponderance of large insertions over large deletions. Genomic rates of both DNA loss through small deletions and DNA gain through large insertions grow with genome size, but at different rates. Whereas the rate of DNA loss grows linearly, the DNA gain grows slower, leading to the existence of a stable equilibrium genome size. Surprisingly, limited empirical evidence suggests that differences in genome size may be driven largely by changes in the per nucleotide rate of DNA loss through small indels.

What is important is that on this view the functionality of nongenic DNA is no longer the primary issue. It is entirely possible that nongenic DNA has a function at any given time, while the reason for the existence of this nongenic DNA has very little to do with this function. Future research into the many factors affecting evolution of genome size should establish the validity and generality of this model.

Appendix

Let us consider a monotonically declining fitness function $f(G)$ that results in a coefficient $s(G) = df/dG$ ($s(G) < 0$), for all values of G. Let N be the absolute size of the population, N_e its effective size, i and d the rates of insertions and deletions, respectively.

Per unit time, in an unconstrained part of the genome, there will be created $2N_i$ insertions that will fix with the probability of $2s_i/(1 - \exp(-4N_e s_i))$ (with $s_i < 0$). There will also be $2N_d$ deletions fixing with the probability of $2s_d$ ($s_d > 0$). We assume here that N_e is large and s_i and s_d are small. Because deletions and insertions are always much smaller than genome size, selection coefficients for both should be proportional to their length (Petrov and Hartl, 2000). If the absolute value of the coefficient of proportionality is α, then $s_i = -\alpha L_i$ and $s_d = \alpha L_d$, where L_i and L_d stand for the average size of insertions and deletions, respectively. At equilibrium the total length of fixed deletions should be equal to that of fixed insertions, giving us the following condition:

$$-\alpha(L_i)^2 i/(1 - \exp(\alpha 4N_e L_i)) = d\alpha(L_d)^2.$$

After simplification assuming that selection is strong ($|4N_e s_i| \gg 1$), we obtain

$$iL_i^2/dL_d^2 = \exp(\alpha 4N_e L_i).$$

Thus if at a particular genome size, selection coefficient per indel of 1 bp in length (α) is related to ratio of the deletion and insertion rates and sizes in this way, the equilibrium will be attained. The stability of this equilibrium depends on the shape of the fitness function. Under the assumptions of the junk DNA theory, the fitness function is a declining concave function of genome size and $iL_i > dL_d$. In such a case the equilibrium will be stable.

However, we still need to explain how this model could produce the power function relationship between the rate of DNA loss through small indels and genome size (Fig. 4 and Formula (1)). This is where we run into a difficulty. In Fig. 4 and Formula (1), practically all of the difference in the rate of DNA loss is due to the differences in the sizes and rates of deletions. We can thus substitute dL_d with $kG^{-1.3}$ using our empirical findings (1). We then find that selection coefficient per 1 bp indel (α) has to be related to genome size in the following way:

$$iL_i^2/kG^{-1.3}L_d = \exp(\alpha 4N_e L_i).$$

and finally

$$\alpha = 1/(4N_e L_i)(\ln(iL_i^2) - \ln(kG^{-1.3}L_d).$$

It is unclear why we would expect strength of selection to depend on all of these parameters across the studied organisms.

ACKNOWLEDGMENTS

I am very grateful to Aaron Hirsh, Douda Bensasson, Jerel Davis, Emile Zuckerkandl, Marc Feldman, Luciano Brocchieri, and an anonymous reviewer for extremely helpful and insightful comments. I also thank the editors, Samuel Karlin and Allan Campbell, for the opportunity to contribute to this issue of TPB.

REFERENCES

Albertini, A. M., Hofer, M. *et al.* 1982. On the formation of spontaneous deletions: The importance of short sequence homologies in the generation of large deletions, *Cell* **29**(2), 319–328.

Bebenek, K., and Kunkel, T. A. 1990. Frameshift errors initiated by nucleotide misincorporation, *Proc. Natl. Acad. Sci. USA* **87**(13), 4946–4950.

Bensasson, D., Petrov, D. A. *et al.* 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers, *Mol. Biol. Evol.* **18**(2), 246–253.

Bensasson, D., Zhang, D. *et al.* 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses, *Trends in Ecology & Evolution* **16**(6), 314–322.

Bergman, C. M., and Kreitman, M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences, *Genome Res.* **11**(8), 1335–1345.

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

Charlesworth, B., and Langley, C. H. 1989. The population genetics of Drosophila transposable elements, *Ann. Rev. Genet.* **23**, 251–287.

Doolittle, W. F., and Sapienza, C. 1980. Selfish genes, the phenotype paradigm and genome evolution, *Nature* **284**, 601–603.

Gojobori, T., Li, W. H. *et al.* 1982. Patterns of nucleotide substitution in pseudogenes and functional genes, *J. Mol. Evol.* **18**(5), 360–369.

Graur, D., Shuali, Y. *et al.* 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans, *J. Mol. Evol.* **28**(4), 279–285.

Gregory, T. R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma, *Biol. Rev. Camb. Philos. Soc.* **76**(1), 65–101.

Gregory, T. R., and Hebert, P. D. 1999. The modulation of DNA content: Proximate causes and ultimate consequences, *Genome Res.* **9**(4), 317–324.

Gu, X., and Li, W.-H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment, *J. Mol. Evol.* **40**, 464–473.

Harrison, P. M., Echols, N. *et al.* 2001. Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome, *Nucleic Acids Res.* **29**(3), 818–830.

Hartl, D. L., Lohe, A. R. *et al.* 1997. Regulation of the transposable element mariner, *Genetica* **100**(1–3), 177–184.

Kidwell, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes, *Genetica* (in press).

Kidwell, M. G., Kidwell, J. F. *et al.* 1977. Hybrid dysgenesis in *Drosophila*: a syndrome of aberrant traits including mutation, sterility, and male recombination, *Genetics* **86**, 813–833.

Kunkel, T. A. 1986. Frameshift mutagenesis by eucaryotic DNA polymerases in vitro, *J. Biol. Chem.* **261**(29), 13581–13587.

Kunkel, T. A. 1990. Misalignment-mediated DNA synthesis errors, *Biochemistry* **29**(35), 8003–8011.

Li, W. H., Gojobori, T. *et al.* 1981. Pseudogenes as a paradigm of neutral evolution, *Nature* **292**(5820), 237–239.

Ling, H., Boudsocq, F. *et al.* 2001. Crystal structure of a Y-family DNA polymerase in action: A mechanism for error-prone and lesion-bypass replication, *Cell* **107**(1), 91–102.

Lozovskaya, E. R., Nurminsky, D. I. *et al.* 1999. Genome size as a mutation–selection–drift process, *Genes Genet. Syst.* **74**(5), 201–207.

Malik, H. S., Burke, W. D. *et al.* 1999. The age and evolution of non-LTR retrotransposable elements, *Mol. Biol. Evol.* **16**(6), 793–805.

Nuzhdin, S. V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number, *Genetica* **107**(1–3), 129–137.

Ohno, S. 1972. So much "junk" in our genomes. *Evolution of Genetic Systems*, *Brookhaven Symp. Biol.*, pp. 366–370. H. H. Smith.

Ophir, R., and Graur, D. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids, *Gene* **205**(1–2), 191–202.

Orgel, L. E., and Crick, F. H. C. 1980. Selfish DNA: The ultimate parasite, *Nature* **284**, 604–607.

Petrov, D. A. 1997. Slow but steady: Reduction of genome size through biased mutation, *The Plant Cell* **9**, 1900–1901.

Petrov, D. A. 2001. Evolution of genome size: new approaches to an old problem, *Trends Genet.* **17**(1), 23–28.

Petrov, D. A. 2002. DNA loss and evolution of genome size in Drosophila, *Genetica* (in press).

Petrov, D. A., Chao, Y.-C. *et al.* 1998. Pseudogene evolution in Drosophila suggests a high rate of DNA loss, *Mol. Biol. Evol.* **15**, 1562–1567.

Petrov, D. A., and Hartl, D. L. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups, *Mol. Biol. Evol.* **15**(3), 293–302.

Petrov, D. A., and Hartl, D. L. 2000. Pseudogene evolution and natural selection for a compact genome, *J. Hered.* **91**(3), 221–227.

Petrov, D. A., Lozovskaya, E. R. *et al.* 1996. High intrinsic rate of DNA loss in Drosophila [see comments], *Nature* **384**(6607), 346–349.

Petrov, D. A., Sangster, T. A. *et al.* 2000. Evidence for DNA loss as a determinant of genome size, *Science* 287(5455), 1060–1062.

Petrov, D. A., Schutzman, J. L. *et al.* 1995. Diverse transposable elements are mobilized in hybrid dysgenesis in *Drosophila virilis*, *Proc. Natl. Acad. Sci. USA* **92**, 8050–8054.

Robertson, H. M. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss, *Genome Res.* **8**(5), 449–463.

Robertson, H. M. 2000. The large SRH family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses, *Genome Res.* **10**(2), 192–203.

Robertson, H. M., and Martos, R. 1997. Molecular evolution of the second ancient human mariner transposon, Hsmar2, illustrates patterns of neutral evolution in the human genome lineage, *Gene* **205**(1–2), 219–228.

Robin, G. C., Russell, R. J. *et al*. 2000. The evolution of an alpha-esterase pseudogene inactivated in the *Drosophila melanogaster* lineage, *Mol. Biol. Evol*. **17**(4), 563–575.

SanMiguel, P., Gaut, B. S. *et al*. 1998. The paleontology of intergene retrotransposons of maize, *Nat. Genet*. **20**(1), 43–45.

SanMiguel, P., Tikhonov, A., *et al*. 1996. Nested retrotransposons in the intergenic regions of the maize genome, *Science* **274**(5288), 765–768.

Shabalina, S. A., and Kondrashov, A. S. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes, *Genet. Res*. **74**(1), 23–30.

Shabalina, S. A., Ogurtsov, A. Y. *et al*. 2001. Selective constraint in intergenic regions of human and mouse genomes, *Trends Genet*. **17**(7), 373–376.

Vinogradov, A. E. 1999. Intron–genome size relationship on a large evolutionary scale, *J. Mol. Evol*. **49**(3), 376–384.

Zuckerkandl, E. 2002. Why are large pluralities of noncoding nucleotides required for function in eukaryotes? The impact of epigenetic control on genome size, *Genetica* (in press).