Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 324 (2004) 15–34
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Review

Insertion–deletion biases and the evolution of genome size

T. Ryan Gregory

Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Received 7 July 2003; received in revised form 29 August 2003; accepted 25 September 2003

Received by A.J. van Wijnen

Abstract

Numerous theories have been proposed to account for the pronounced differences in the quantity of non-coding DNA among eukaryotic genomes, but the current repertoire remains incomplete because the only explicit mechanisms it provides involve DNA gain. It has been proposed more recently that biases in spontaneous insertions and deletions (indels) can lead to genome shrinkage by mutational mechanisms alone. The present article provides the first detailed critical discussion of this approach, and covers three different ideas related to it: (1) the general notion of DNA loss by deletion bias, (2) the "DNA loss hypothesis" which supposes that variation in genome size can be attributed to differences in DNA loss rate, and (3) the "mutational equilibrium model" which attempts to describe the long-term evolution of genome size. The mutational equilibrium model is found to be problematic, and it is noted that DNA loss by small indels is too slow in real time to determine variation in genome size above a relatively low threshold. Some alternative explanations for the observed patterns are provided, and the critique also identifies some potential problems with the current dataset. These include a failure to cite a more detailed (and somewhat contradictory) mammalian dataset, a questionable use of arithmetic means with highly skewed data, and important discrepancies among the particular DNA sequences so far analyzed. Overall, evolutionary reductions in genome size are considered important, but the specific mechanism relating to small deletion bias is far too weak to be accepted as a primary determinant of genome size variation in general.

Keywords: C-value; DNA content; DNA loss; Indel; Junk DNA; Pseudogenes; Selfish DNA; Transposable elements

1. Introduction

1.1. The enigma of genome size variation

There is no question that whole-scale DNA sequencing has provided important new insights into the evolution of the genome. Not least among these is an affirmation of the sheer complexity of genome evolution as a historical process. As prime examples, one need only consider the strikingly low number of genes and the remnants of pronounced transposable element (TE) activity revealed by the completion of the draft human genome sequence (International Human Genome Sequencing Consortium, 2001). More recent comparisons with other model vertebrate genomes have only heightened the realization that genome evolution is not simply, or even primarily, a story of gene evolution (Aparicio et al., 2002; Hedges and Kumar, 2002; Mouse Genome Sequencing Consortium, 2002).

This intriguing new knowledge aside, the fact that the history of genome evolution is not dominated by changes in the characteristics of coding sequences has been known in general terms for over half a century. Thus, the lack of correspondence between gene number and organismal complexity, labeled as the "G-value [or N-value] paradox" (Claverie, 2000; Betrán and Long, 2002), is only the modern equivalent of the disconnect between genome size and morphological complexity first observed in the late 1940s and dubbed the "C-value paradox" in the early 1970s (Thomas, 1971). But as puzzling as these observations are, they are only "paradoxical" in the light of erroneous assumptions about genome biology. The C-value paradox, for example, was solved with the discovery that most DNA is non-coding.

The solution to the C-value paradox raised numerous new questions about genome evolution, just as the solution

Abbreviations: bp, base pair; DOA, dead-on-arrival; Indel, insertion or deletion; LTR, long terminal repeat; Mb, mega (million) base pairs; numt, nuclear pseudogene of mitochondrial origin; pg, picograms (10^{-12} g, or ~980 Mb); TE, transposable elements; UV, ultraviolet light.

E-mail address: rgregory@genomesize.com (T.R. Gregory).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2003.09.030

to the G-value paradox will undoubtedly engender new puzzles regarding development and gene regulation. The many questions relating to the reasons for the highly divergent levels of acquisition, maintenance, and loss of non-coding DNA among eukaryotes, and the biological significance thereof, make up the much larger "C-value enigma" (Gregory, 2001a).

1.2. DNA loss: a missing piece of the puzzle

Various approaches have been taken to the C-value enigma, the most famous being the "selfish DNA" and "junk DNA" theories. Though these terms are often misused, they have specific meanings relating to the proposed mechanisms of non-coding DNA accumulation (reviewed in Gregory, 2001a). Selfish DNA is that which is spread by the actions of egoistic sequences like transposable elements (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), while junk DNA refers to the accretion of defunct gene duplicates (Ohno, 1972), or "pseudogenes". There are also two prominent "optimal DNA" theories which focus on a different aspect of the C-value enigma, namely the consequences (but not the mechanisms) of DNA content change. These are the "nucleoskeletal" and "nucleotypic" theories which, though differing substantially in their specifics, both describe genome size variation as the outcome of selection via the intermediate of cell size (see Cavalier-Smith, 1985; Gregory, 2001a for reviews).

The main difficulty with this current repertoire of genome size theories is that it provides no explicit mechanism for DNA loss. Both mutation pressure theories describe a unidirectional force for genomic growth, and while the nucleotypic theory may often postulate the necessity of genome shrinkage, it has little to say regarding the mechanisms by which this might be accomplished.

Reductions in genome size have been postulated, or indeed identified, in several plant and animal groups (e.g., Jockusch, 1997; Watanabe et al., 1999; Wendel et al., 2002a). In some cases, selection for decreased DNA content would be consistent with known relationships between genome size and cellular/organismal features such as metabolic rate, development time, or body size, but selective hypotheses are not necessarily accurate, nor do they provide a mechanistic explanation for DNA loss. More recently, an approach to genome size evolution has emerged that seeks to avoid both of these problems by providing a non-selective, mechanistic explanation for DNA loss (Petrov, 2001).

Specifically, a causal link is proposed to exist between genome size and the rate at which DNA is removed from the genome by the accumulation of small deletions that are not compensated for by insertions. Small genomes, in particular, have been interpreted as resulting from this neutral mutational mechanism of biased insertions and deletions ("indels"), and not (exclusively) as a product of nucleotypic selection. Thus, Petrov (2002a) argues that "differences in genome size may be driven largely by changes in the per nucleotide rate of DNA loss through small indels".

Obviously, this is an idea worthy of serious consideration. Yet, to date only one brief discussion of the merits of the small indel bias approach has been made available (Gregory, 2003a). The present article provides the first detailed critical appraisal of this approach to genome size evolution and its likely relevance to the study of the C-value enigma.

2. Indel bias, the DNA loss hypothesis, and the mutational equilibrium model

That genomes can decrease in size through evolutionary time is not an especially controversial notion and is not at issue here. Instead, the question under consideration relates to a specific mechanism of DNA loss—to wit, whether DNA is often lost, in biologically relevant quantities, by biases in *small indels* (i.e., <400 bp). The models of genome size change to be evaluated here deal exclusively with such small indels, and neither preclude nor refute the importance of large gains and losses of DNA.

There are, in fact, three separate issues involved in this discussion. First among these is the simple premise that indel biases can lead to significant losses of DNA. Second is the "DNA loss hypothesis", that such a mechanism has played an important role in shaping genome size diversity among eukaryotes. Third, the "mutational equilibrium model" of genome size evolution, which is based on the DNA loss hypothesis but makes many additional claims about the process of genome size evolution. It is important to note that although there has been a logical development from the first concept to the third, they need not be accepted in toto.

2.1. DNA loss by indel bias

It has long been recognized that small deletions outnumber insertions of similar size in protein-coding sequences (e.g., de Jong and Rydén, 1981), and larger and more frequent deletions have since been reported many times for non-coding regions as well (see Petrov, 2002a). According to Petrov (2002a), this deletion bias is to be expected on the basis of the thermodynamics of replication slippage, in which an insertion requires the melting and rereplication of a segment of previously duplicated DNA, whereas deletions involve only a skipping of unreplicated bases. While true in theory, this seems unlikely to provide a complete explanation because indels larger than 5 bp (or so) are probably generated by unequal crossing over rather than replication slippage (see below), thereby removing this thermodynamic disparity between insertions and deletions. In any case, the empirical demonstration of the indel bias in non-coding DNA is sufficient, even in the absence of a convincing explanation, to indicate that the outcome of many small indel events would indeed be a net loss of DNA.

That said, it is crucial to bear in mind that this proposed mechanism of DNA loss does not relate to any insertions or deletions outside the stated 1–400 bp range, so this is necessarily relevant to only a subset of possible mechanisms of DNA content change. In terms of large indels, the tendency will be towards DNA gain, because while large deletions are unlikely due to their effects on gene function, only the location, and not the length, of insertions will determine their impacts on genes. So, although genomes may indeed shrink by a slow process of deletion bias in small indels, they may also be expected to grow by a faster mechanism of insertion bias in large indels. For example, transposable elements (which make up a large fraction of most eukaryotic genomes) are larger than 400 bp, such that each proliferation event will necessarily fall outside the range of small indels being discussed. Thus, even accepting the logic and empirical support for DNA loss by deletion bias in small indels, it is important to note that this is only one of several mechanisms of genome size change.

2.2. The DNA loss hypothesis

The first detailed examination of indels in segments of non-coding DNA was performed by Graur et al. (1989). By comparing the relative rates and sizes of indels in a sample of 22 processed pseudogenes from human, 14 from mouse, and 16 from rat, Graur et al. (1989) were able to determine that deletions outnumbered insertions, and that rodent pseudogenes appeared to lose DNA less slowly than those of humans. This observation could not be extended to the most common genetic model of all until some time later, however, because the genome of *Drosophila* contains few pseudogenes (Jeffs and Ashburner, 1991; but see Harrison et al., 2003). This difficulty was eventually circumvented when Petrov et al. (1996) used non-LTR transposable elements (*Helena*) as pseudo-pseudogenes. These elements are self-replicating when active, but upon insertion they commonly experience a truncation at the 5′ end which renders them "dead-on-arrival (DOA)" and therefore incapable of subsequent transcription. Once inactivated, they are assumed to evolve neutrally in the same manner as true pseudogenes.

The details of the dead-on-arrival retrotransposon study have been given many times elsewhere (Petrov et al., 1996; Petrov and Hartl, 1997, 2000; Lozovskaya et al., 1999; Petrov, 2001, 2002b), and need not be repeated here. Suffice it to say that deletions were larger and more frequent than insertions, and that the overall relative rate of DNA loss—calculated as the average number of base pairs lost per base pair (bp) substitution—was much higher in *Drosophila* than the value calculated for mammals based on the data of Graur et al. (1989). It was therefore suggested that "mutation pressure in the form of spontaneous deletions of DNA may account for the reduction in genome size that ultimately leads to a compact genome such as that found in *Drosophila*" (Lozovskaya et al., 1999).

To further test this idea, Petrov et al. (2000) and Bensasson et al. (2001) examined the indel spectra of two more insects, chosen to exceed the *Drosophila* C-value [0.18 picograms (pg)] by order of magnitude intervals: the Hawaiian cricket (*Laupala cerasina*) at 1.9 pg, and the mountain grasshopper *Podisma pedestris* at 16.9 pg. (That orthopterans were chosen for this comparison with *Drosophila* was something of a necessity, since no other order of insects displays anything like this level of variation; Gregory, 2001b.) In *Laupala*, the "pseudogene" of choice was another dead-on-arrival non-LTR transposon (*Lau1*), while in *Podisma* it was nuclear pseudogenes of mitochondrial origin ("numts"). Thus, the type of DNA element examined differs substantially among the species studied to date. It remains to be seen whether the assumption that all types of quasi-pseudogene sequences evolve in the same way will withstand further examination (Petrov, 2002b), but if so then this will provide good evidence for the universality of key genome-level evolutionary processes.

The results of the two orthopteran studies revealed an intriguing relationship between genome size and relative DNA loss rate. *Podisma*, with the largest genome, loses DNA more slowly than mammals, which in turn have slower DNA loss rates than *Laupala*. Of course, *Drosophila*'s minuscule genome has the fastest DNA deletion pattern of the lot. More recently, pufferfishes have also been found to show DNA loss rates roughly in proportion to their genome sizes (Neafsey and Palumbi, 2003). That is to say, there is an apparent correlation between genome size and relative DNA loss rate measured as bases lost per base pair substitution. (Relative DNA loss rate is calculated as the DNA lost by deletions [average deletion size × ratio of deletions to substitutions] minus DNA added by insertions [average insertion size × ratio of insertions to substitutions].) The overall difference in relative DNA loss rate is about 50-fold, corresponding to a roughly 100-fold range in genome sizes.

At present, the relationship is based on only five data points, although analyses are ongoing. More recently, Petrov (2002a) has included *Caenorhabditis elegans* in the expected location on the DNA loss-genome size plot, although this value should be interpreted with caution because it was based on rough calculations rather than direct measurements. It does bear mentioning that the force of deletion bias would need to be a particularly powerful in nematodes since *Caenorhabditis* undergoes more frequent duplications, pseudogene formations, and intron gains than *Drosophila* (Semple and Wolfe, 1999; Robertson, 2000; Friedman and Hughes, 2001; Cavalcanti et al., 2003). And while *C. elegans* does indeed undergo large deletions (up to nearly 800 bp—twice as large as in *Drosophila*), it also displays even larger insertions (up to almost 4700 bp) (Robertson, 2000). It is also clear that *C. elegans* possesses more pseudogenes than *Drosophila* (Harrison et al., 2003), and there may in fact be reason to expect an overall *increase* in genome size by indels in *C. elegans*, despite its tiny

genome (Comeron, 2001). For these reasons, *C. elegans* will not be included in the following discussions of the existing DNA loss rate data. The nature and significance of the correlation, its overall relevance to genome size evolution, and some alternative explanations to account for it (i.e., besides assuming that small indel bias determines genome size) will be discussed in some detail in later sections.

2.3. The mutational equilibrium model

Two very different implications of the DNA loss hypothesis have been envisaged by its proponents. On the one hand, Hartl (2000) suggests that "the realized genome size at any time may result from a dynamic and constantly shifting balance between slow deletion and rapid transposon proliferation". In contrast, Petrov (2002a) has developed a "mutational equilibrium model" which postulates that "eventually, at the stable equilibrium genome size value, rates of growth and loss equal each other". Only the second of these is potentially controversial, because it is explicitly dependent on the DNA loss hypothesis and assumes that small indel bias is the predominant force in shaping long-term trends in genome size evolution.

Under the mutational equilibrium model, the activities of transposable elements, large duplications, and other such "rapid" mechanisms of genome size change appear only as noise around a primary signal of slow DNA loss. In particular, genome size is seen in mutational terms as a balance between DNA gain by large insertions and slow but steady DNA loss by small deletions. Natural selection, as a force acting to limit excessive genome size growth (mutation pressure theories) or to shape genome sizes adaptively (optimal DNA theories), is not precluded by such a model, but its relevance, if any, is considered only secondary.

According to Petrov's (2002a) model, the "rate of DNA loss through small deletions scales linearly with genome size, whereas DNA gain through large insertions scales slower than linear" (see his Fig. 5). As a result, "different rates of DNA loss per nucleotide lead to different equilibrium genome sizes". However, this assertion is clearly contradicted by the existing data, given that the correlation between DNA loss rate and genome size is only significant when log-transformed, and is therefore decidedly nonlinear (Petrov, 2002b). Indeed, Petrov (2002a) reports that DNA loss rate and genome size scale as a 1.3 power function.

In fact, this nonlinear scaling plays a prominent role in Petrov's (2002a) arguments against traditional genome size theories, in which he points out that neither the junk DNA nor adaptive theories predict this specific 1.3 power function. At least three objections can be raised against this argument. First, one may note that *of course* the junk DNA and adaptive theories do not predict a 1.3 power function relationship between the rate of DNA loss per base pair substitution and genome size. This is because they do not predict (or deny) *any* particular relationship between these parameters. Likewise, the DNA loss hypothesis does not predict (or deny) a strong positive correlation between genome size and cell size (reviewed in Gregory, 2001a), but this is irrelevant to the validity of the mutational model as assessed within its proper domain. Second, the current mutational equilibrium model provides no mechanistic explanation for why this relationship should take the form it does (and, as discussed above, actually assumes something different). This is a far more serious omission for the DNA loss model than it is for the junk DNA and adaptive theories, for the obvious reason that it forms the basis of the former but has little to do with the latter. Third, with so few data presently available, it is highly possible that the relationship will change considerably with the addition of more taxa. Even the addition of one point for *C. elegans* makes the relationship appear more sigmoidal than linear when log-transformed (see Fig. 4 in Petrov, 2002a). More generally, there may also be statistical reasons to doubt the DNA loss rates as currently calculated for the pivotal organisms (see below).

A second scaling function outlined in the mutational equilibrium model is a 1/4 power function between insertion rates and genome size (Petrov, 2002a). This relationship is not based on an empirical result, but is instead a necessary assumption of the model in order to explain the 1.3 power scaling of DNA loss rate versus genome size. In light of the deleterious effects of insertions on gene sequences, one might expect that insertions (e.g., by transposable elements, tandem duplications, etc.) should be *more* common in larger genomes with more non-coding DNA to serve as "safe" insertion sites. The proposed escape from this logical expectation takes two forms. First, it is pointed out that the insertion rates of transposable elements must not increase faster than linear, since TE copy number does not increase exponentially to astronomical levels (Petrov, 2002b). However, at best this argument establishes that TE insertion rates increase no faster than linear with genome size; it says nothing about a 1/4 power function. More importantly, the neglected point is that this limitation is probably imposed by selection against such uncontrolled spread, which is what both mutation pressure and nucleotypic theories would expect, but which the mutational equilibrium model attempts to avoid. Second, once genomes grow to a certain size, large deletions that do not affect genic regions will become possible and will counteract these large insertions (Petrov, 2002b). But again, this does not specifically predict a 1/4 power function, but only a relationship that scales less than or equal to linear. More importantly, this explanation does not rescue the mutational equilibrium model at all, since it clearly implies that the balance of *large* indels, not small ones, ultimately plays the dominant role in shaping genome size.

2.4. DNA loss by indel bias: a summary

To review, the most straightforward aspect of DNA loss is the notion that the operation of small indels should be biased towards deletion and therefore tend to produce a net

loss of DNA. From this emerged the *DNA loss hypothesis*, which proposes that variation in genome size is largely the product of differences in the relative rates of DNA loss by small indel bias. By extrapolation from the DNA loss hypothesis has developed a more explicit *mutational equilibrium model*, which provides a description of the way in which long-term genome size evolution is seen to proceed. It is important to recognize that these ideas have led from one to the next in both logical and temporal sequence, but that they nonetheless remain separate postulates. One can accept the reality of indel biases without accepting the DNA loss hypothesis (e.g., genome size and indel rates may be correlated, but the latter does not necessarily determine the former) and one can similarly accept the DNA loss hypothesis while still rejecting the mutational equilibrium model (e.g., indel biases may exert an influence on genome size, but other factors are much more important in determining the patterns of variation among species).

It should be clear from the above discussion that the mutational equilibrium model suffers from some important shortcomings. However, even if one is wisely hesitant to accept the current formulation of the model, it should still be considered worthy of further development and empirical testing. Pending these improvements, the present discussion will focus on the concept of indels in influencing DNA contents, and especially on the attempt of the DNA loss hypothesis to link this in a strongly causal way to the profound variation in genome size observed among taxa.

3. Bacteria and barley: examples of deletion bias at work

Although the DNA loss hypothesis is based on data from animals, the two best examples of indel mechanisms at work come from bacteria and grasses. These cases illustrate that such a process of indel biases can be important for the evolution of genome size, even though differences in the underlying mechanisms mean that they lend no direct support to the DNA loss hypothesis per se.

3.1. Genome shrinkage in endosymbiotic bacteria (and algae?)

Nucleotypic effects of bulk DNA are often considered to be particularly relevant in single-celled organisms, for obvious reasons. For example, among protists there is a positive correlation between DNA content and cell size, and a negative relationship between these and division rate (e.g., Shuter et al., 1983; Wickham and Lynn, 1990). It has largely been assumed that similar relationships would exist among bacteria, particularly given that their single circular chromosome contains only one replicon origin, but a recent comparison of doubling times under laboratory conditions revealed no such correlation with genome size among prokaryotes (Mira et al., 2001). Bacteria also differ from protists in that they show no evidence of the old "C-value paradox", since there is a direct correlation between genome size and gene number in these organisms (Mira et al., 2001).

It has been noted several times that free-living bacteria have larger genomes than those dependent on multicellular hosts. Selection for rapid replication among parasites is a commonly cited explanation, but the lack of association between doubling time and genome size calls this into question (Mira et al., 2001). While there is some reason to believe that free-living bacteria may be more prone to genome expansion than parasitic forms (Stepkowski and Legocki, 2001; Liò, 2002), deletion bias may be the primary determinant of this pattern. Specifically, following a shift in lifestyle from free-living to obligate endosymbiosis or parasitism, many previously essential genes lose their functional relevance (Mira et al., 2001; Frank et al., 2002), and as a result these obsolete genes are eventually lost as they succumb to a mutational bias towards small deletions (Mira et al., 2001).

Again, this bacterial example shows that deletion bias can indeed shape genome size diversity in some organisms, but it does not lend support to the DNA loss hypothesis. First, note that the input of deletion bias to patterns of bacterial genome size variation involves the loss of existing genes that are suddenly no longer maintained by selection. It does not relate directly to elements like pseudogenes and transposable elements that are continually added to the genome. Second, while deletion bias may account for much of the relatively minor genome size variation found among free-living versus parasitic bacteria, it does not explain why prokaryotes in general have genomes much smaller than those of almost all eukaryotes (the exception being a few parasitic protists with tiny C-values), or why non-coding DNA is so sparse in all bacterial genomes. It therefore contributes very little to the understanding of the C-value enigma. In reality, large DNA additions of the type that make some eukaryotic genomes so big probably *are* selected against in bacteria on the basis of genome size-related effects on division rate. Doubling time may not correlate with C-value over the tiny range in genome sizes available for analysis, but how quickly could a bacterium with a human-sized genome divide, especially with only one replicon origin?

It is now well established that organelles such as mitochrondria emerged by a process of primary endosymbiosis, in which a formerly free-living bacterium was ingested and retained by another cell. In the case of mitochondria, which exist in multiple copies and replicate independently in the cell, there is still good reason to believe that genome sizes were reduced in response to selection (Mira et al., 2001; Selosse et al., 2001). A profound reduction in genome size has also characterized the evolution of secondary endosymbionts, which are represented by "nucleomorphs"—greatly reduced former nuclei found along with normal nuclei within the cells of some algae. It has been suggested in this case that the extreme genome size reduction in the

nucleomorph is not the result of selection, but can instead be explained by mutational mechanisms of DNA loss (Gilson, 2001; Gilson and McFadden, 2002). In a sense, the question is therefore whether these former algae more closely resemble endosymbiotic bacteria or organelles in terms of genome size shrinkage mechanisms.

The distinguishing feature of nucleomorphs is that they are former eukaryotic nuclei, not bacteria. As such, their genomes initially contained not only functional genes, but also non-coding sequences of various descriptions. Therefore, while the loss of function could certainly have allowed genes to be removed by deletion bias as in endosymbiotic bacteria, there is necessarily more to the story. Similarly, transfer of genes to the primary nucleus would account for some DNA loss, and this may or may not be associated with selection as in organelles. Since nucleomorphs exist in only one copy per cell and are therefore not competing with one another for rapid intracellular replication, the form of selection would necessarily be different in any case (Gregory, 2001a).

But why should non-coding sequences, especially actively replenishing ones like transposable elements, be lost by deletion bias after secondary endosymbiosis? According to Gilson (2001), the loss of sexual reproduction and the associated process of meiotic recombination would have prevented transposable elements from being replaced, thereby making them susceptible to gradual removal by deletion bias. In support of this hypothesis, Gilson (2001) cites the example of sexual versus asexual rotifers, the latter of which lack an entire class of transposable elements that is found in abundance in the former (Arkhipova and Meselson, 2000). However, although sexual reproduction is relevant to the spread of *retrotransposons* in particular, this is not true of all transposable elements. For example, in the long-asexual bdelloid rotifers, retrotransposons may be absent, but *mariner*-like elements are not (Arkhipova and Meselson, 2000). This is because *mariner* elements spread by horizontal transfer, and therefore do not require sex to get around. Moreover, Arkhipova and Meselson (2000) argue that a loss of sex allows retrotransposons to be lost by several mechanisms, including both deletion bias *and* selection for their elimination. It also bears noting that although genome sizes are known for only a few rotifer species, the asexual bdelloids appear to have *larger* genomes than their sexually reproducing monogont relatives (Mark Welch and Meselson, 1998, 2003).

The extreme reduction in genome size in nucleomorphs—to the point of having overlapping gene sequences—also suggests something more powerful than simple deletion bias in the absence of retrotransposon proliferation (Cavalier-Smith, 2002). Indeed, the example of nucleomorphs has been used as evidence in favour of at least one optimal DNA theory (Beaton and Cavalier-Smith, 1999; Cavalier-Smith and Beaton, 1999).

In summary, there is evidence that deletion bias affects genome size in endosymbiotic bacteria, relating specifically to the loss of genes. Deletion bias does not seem to apply to other endosymbionts like organelles or nucleomorphs, however.

3.2. LTR transposons in barley (and other plants)

In 1997, Bennetzen and Kellogg posed the question, "Do plants have a one-way ticket to genomic obesity?". While various modes of genome size increase are well known to operate in plants, no process for DNA loss had yet been described. Nevertheless, they suggested that such a DNA loss mechanism could operate in plants, at least in principle. Petrov (1997) countered that the deletion bias mechanism described in *Drosophila* could provide a "return ticket" for plants and "significantly reduce genome size over evolutionary timescales". Indeed, a deletional mechanism capable of at least limiting genome growth has since been found to operate in barley and its relatives (*Hordeum* spp.) (Vicient et al., 1999a,b; Rabinowicz, 2000). As with the bacteria discussed above, this mechanism demonstrates the importance of DNA loss as a general process but differs crucially from that underlying the DNA loss hypothesis.

In contrast to the non-LTR *Helena* elements studied in *Drosophila*, the DNA sequence in question here is the LTR retrotransposon *BARE-1*. By definition, these elements are endowed with long terminal repeats (LTRs) which, once inserted, tend to promote recombination between homologous LTRs. This recombination frequently results in the loss of one of the LTRs and the internal domain necessary for transposition. The outcome of repeated recombinational deletions of this sort has been a 40-fold excess of "solo LTRs" relative to internal domains of the *BARE-1* element (Vicient et al., 1999a,b; Rabinowicz, 2000). The phenomenon is not restricted to this particular element since other LTR transposons show similar patterns in *Hordeum* (Shirasu et al., 2000).

The deletions involved in this case are consistently large (i.e., most of the *BARE-1* element), not variable small to medium-sized ones. This would therefore represent a much stronger deletional mechanism than that observed with *Helena* elements in *Drosophila*. This deletion bias in *Hordeum* therefore probably acts to limit the expansion of the genome by transposition much more effectively than the small deletion bias mechanism in *Drosophila*. However, there is a crucial difference between the indel bias mechanisms of the two groups: since the recombination rate in barley is additive with the presence of LTRs, the more elements that are present, the more powerful is the recombinational loss mechanism. This means that larger genomes probably undergo this process more frequently than small ones, and that DNA loss rate and genome size would be *positively* correlated in this case. Moreover, this would also represent an example of DNA loss rate being influenced by genome size, and not the reverse as assumed under the DNA loss hypothesis. It is also informative in this regard that while the DNA loss hypothesis would predict that related

plants with smaller genomes would lose DNA more quickly than barley, it is obvious from comparisons of *Hordeum* and *Zea* that the smaller genome of maize undergoes far *fewer* of these recombinational deletions (SanMiguel et al., 1998; Vicient et al., 1999a,b; Rabinowicz, 2000; Shirasu et al., 2000; García-Martínez and Martínez-Izquierdo, 2003).

The situation in barley also allows a comparison of the traditional selection-based approaches to genome size evolution and the mutational view favoured by DNA loss proponents. In a recent study of *Hordeum spontaneum* from different microclimatic regions of Evolution Canyon in Israel, Kalendar et al. (2000) found that plants living on higher regions of the slopes had more copies of *BARE-1* elements than those at lower elevations. Two interpretations are possible for this observation. On the one hand, this may be due to selection for larger genome sizes in more xeric regions (Vicient et al., 1999a,b; Kalendar et al., 2000). This would be consistent with other correlations reported between environmental conditions and genome size in *Hordeum* and its relatives (e.g., Rayburn and Auger, 1990; Bullock and Rayburn, 1991; Poggio et al., 1998; Turpeinen et al., 1999). Alternatively, it might be argued that the increased TE activity results from a more neutral mutational basis, for example from greater exposure to ultraviolet light (UV) radiation at the top of the slope. Such an effect has been observed with *Mutator* transposons in maize (Walbot, 1999), although it does bear noting that while the two slopes differ in their exposure to sunlight, there is little effect of elevation on UV exposure within either slope (Schulman, personal communication). In any case, it is obvious that these hypotheses would be very difficult to disentangle based only on the correlation between altitude and *BARE-1* copy number. However, it also appears that the rate of recombinational loss is lower in high-elevation plants (Kalendar et al., 2000); recall that it is normally the case that more *BARE-1* elements mean faster rates of deletional loss (Vicient et al., 1999a,b). This observation would seem to favour the selective interpretation, since the plants living in more arid conditions are apparently deleting transposons more slowly, and not just acquiring them more quickly, than individuals lower in the canyon. If the selective interpretation holds, then this too would clearly be in conflict with the mutational model of genome size, although not with the general notion of DNA loss by indel biases.

In a recent analysis, Devos et al. (2002) examined the patterns of LTR retroelement loss by unequal crossing over in the well-known small-genomed plant *Arabidopsis thaliana*. Like maize, *Arabidopsis* shows evidence of "a surge of retrotransposon amplification in recent times" (Devos et al., 2002) as well as large-scale duplications and polyploidy (Arabidopsis Genome Initiative, 2000; Simillion et al., 2002; Blanc et al., 2003), but in this case the genome remains very small (more on the maize genome below). Whereas maize LTR elements are mostly intact and solo LTRs are rare (SanMiguel et al., 1998), the ratio of solo LTRs to intact elements in *Arabidopsis* was found to be 1:1 (Devos et al., 2002). Recall that in barley, solo LTRs outnumber intact elements 40:1 (Vicient et al., 1999a,b; Rabinowicz, 2000). In this regard, DNA loss by unequal recombination in *Arabidopsis* is clearly much slower than in barley, despite its much smaller genome.

In terms of other mutational mechanisms, *Arabidopsis* was found to undergo larger and more frequent deletions during double-strand break repair than tobacco, which has a genome about 20 times larger (Kirik et al., 2000; Orel and Puchta, 2003). However, the average deletion sizes in this case fell well outside the stated 400 bp range, and therefore are not of the type described in the DNA loss hypothesis. Moreover, deletions of this size are probably large enough to be visible to selection. It also bears mentioning that *Arabidopsis* is a weed—i.e., a member of an eclectic group of plants for which small genome sizes are typical and selection for rapid development quite plausible (Bennett et al., 1998).

Like the other plants described above, rice (*Oryza sativa*) appears to have undergone a relatively recent amplification of LTR retrotransposons, but a surplus of solo LTRs from several different transposon sequences indicates that some of the added DNA has been subsequently lost by unequal recombination among LTRs (Vicient and Schulman, 2002; Vitte and Panaud, 2003). Again, such a mechanism does appear to play an important role in impeding the spread of LTR retrotransposons within the genome. However, in every one of these cases it must be borne in mind that no matter how fast this mechanism is, it "can never neutralize the genome expansion driven by LTR-retrotransposon amplification because [at least] a solo LTR is retained" (Devos et al., 2002). The options afforded by this mechanism are fast versus slow net growth, and not increase versus decrease per se, depending on the extent to which portions of the inserted DNA are eliminated. And again, the rate at which DNA is lost in this way appears to correlate positively, and not inversely, with genome size.

4. Is DNA loss rate a determinant of genome size?

Having acknowledged that the concept of indels is at least an important one for genome evolution in bacteria and certain plants, the question is raised as to how generally applicable such a mechanism is in explaining the C-value enigma. That is to say, has DNA loss been of major importance in the evolution of most genome sizes, and if so, over what temporal and taxonomic scales?

4.1. From relative to real time

As currently formulated, the DNA loss hypothesis is framed in terms of relative rates of DNA loss. Again, relative loss rate is calculated as the DNA lost by small deletions minus the DNA added by small insertions, with both measured per base pair substitution, not per unit time. To determine the relevance of DNA loss to genome size evolu-

tion in real life, some measure is required of how quickly or slowly DNA will actually be deleted by this mechanism. Obviously, this will vary according to the rate of substitutions per unit time, which itself varies among organisms.

By using estimates of substitution rates to convert to base pairs lost per unit time, and by employing an exponential decay equation, it has been possible to calculate the "half-life" of a DNA element inserted into the genomes of the various organisms so far studied. Thus, a newly inserted "pseudogene" in *Drosophila* has a half-life of about 14 million years. In *Laupala*, it would take closer to 615 million years to delete *half* of a new pseudogene, and in mammals at least 800 million years (Petrov et al., 1996; Petrov, 2002b). Substitution rates are not known for *Podisma*, but the estimated range in pseudogene half-life is between 880 million and 3.5 billion years (Petrov, 2002b). It seems relevant to point out that mammals have been on the Earth for about 160 million years, animals for perhaps 600 million, and life itself for something like 3.8 billion. According to Petrov (2002a), "when we consider the long-term evolution of genome size, over hundreds of million years, slow and persistent indel bias may be just as efficacious as any fast but sporadic force". Yet, except in *Drosophila*, DNA loss by indel bias is an *incredibly* slow process—so slow, one might conclude, as to be irrelevant when placed in the context of absolute units of time.

It would seem that the only difference in genome size that can be attributed, even in part, to DNA loss mechanisms is between *Drosophila* versus all the other taxa studied to date. Differences among crickets, mammals, and grasshoppers must be produced entirely by other factors. The two orthopterans studied suggest that DNA loss is not a prime determinant of genome size variation within orders, and the comparison of mammals with these insects likewise speaks against any general relevance at the phylum level. As will be seen, DNA loss mechanisms based on small indels similarly cannot explain the variation in genome size within the grass family nor within the mammalian class. This means that, based on the current dataset, the relevance of the relationship between genome size and DNA loss rate remains highly ambiguous with regards to the differences apparent among higher taxonomic categories. The remaining level, among species within genera, has not been discussed in detail previously, although it is also possible to evaluate the current dataset from this perspective.

4.2. Grasses and other plants

To date, deletional biases of the type described in *Drosophila* have not been investigated in plants. If one accepts that "there is no reason to believe that plants are different [from insects] in their propensity to lose DNA through spontaneous mutation" (Petrov, 1997), then it should be possible to inquire about the potential relevance of such small-scale mechanisms in grass species. Although it remains smaller than that of barley, the genome of maize is believed to have been doubled by the explosive activity of transposable elements in the last 3 million years (SanMiguel and Bennetzen, 1998). Assuming for the sake of argument that maize loses DNA by deletion bias at the same rate as the cricket *Laupala* (which has a similar genome size), it is possible to consider how long it would take to jettison this new genomic baggage. Again, according to the calculations of Petrov (2002b), and assuming that the transposable elements are now completely inactive, it would take about 615 million years to delete *half* of maize's newly acquired non-coding DNA. Therefore, provided that no more TE expansions, pseudogene duplications, or polyploidization events occur in the meantime, the recent enlargement of the maize genome ought to be mostly undone by deletion bias sometime within the next 1.5 billion years. To put this in context, maize diverged from other teosinte lineages about 15 million years ago, and the grass family Poaceae itself (which also includes wheat, rice, oats, and sorghum) arose around 75 million years ago (Gaut, 2002). Despite the extremely slow rate at which the newly acquired DNA would actually be lost, Petrov (2002a) suggests that the recent doubling of the maize genome is little more than "noise around the long-term equilibrium value". In addition to the temporal difficulties inherent in this equilibrium view, there is also the fact that "rapid change in DNA content, as well as chromosome number, is a hallmark of grass genome evolution" (Gaut, 2002).

The general finding within plant families is that both growth and shrinkage of genome size are common (e.g., Bennetzen, 2002; Wendel et al., 2002a). On the most basic level, it seems likely that the ancestral flowering plant genome was small (Leitch et al., 1998), meaning that shrinkage mechanisms serve primarily to halt the general pattern of increase, or that DNA loss is actually the noise and growth the central signal among plants. In some specific cases, as with maize, it is obvious that indel bias is not a particularly important consideration as compared to factors like transposable element activity, even if these are relatively rare.

4.3. Mammals (and pufferfishes)

It is often pointed out that rodents have both faster pseudogene deletion rates and smaller genomes than humans, although this is usually qualified by reference to how small the absolute differences in loss rate actually are (e.g., Hartl, 2000; Petrov, 2001). Nevertheless, the intended message must be that deletion rates and genome sizes are somehow related in mammals, or else it is difficult to see why this should be mentioned at all.

Indeed, mice and rats do have slightly smaller genomes than humans ($\sim$ 3.2 vs. 3.5 pg). They also have fewer chromosomes ($2n = 40$ for mice, $2n = 42$ for rats). Whether the disparity in DNA content between humans and rodents lies in indel patterns or karyotypic effects is therefore not immediately obvious. However, it is informative to note that the difference in genome size between chimps and humans

is roughly the same (3.75 vs. 3.5 pg), but here the time frame available to change genome size by deletion bias would clearly be prohibitively short. (In addition, note that chimps have more chromosomes than humans, $2n = 48$, and that much of the variation in primate genome sizes is actually due to differential retrotransposon insertion; Liu et al., 2003.)

Cases do exist in which the reverse is true, as with humans versus the common gray squirrel, *Sciurus carolinensis* (Rodentia: Sciuridae), which has a genome about 30% larger than that of humans but only $2n = 40$ chromosomes. Therefore, a test of the indel bias among mammals could readily be conducted without the confounding factor of chromosome number differences, although such a study should not be necessary. This is because in mammals "processed pseudogenes are created at a much faster rate than they are obliterated by the process of pseudogene abridgment [such that] growth of the genome is not significantly retarded by the occurrence of deletions" (Graur et al., 1989). Furthermore, in contrast to *Drosophila*, there is no correlation in mammals between the ages and sizes of individual pseudogenes, suggesting that there is no steady deletion of DNA by mutational mechanisms in this case (Graur et al., 1989; Ophir and Graur, 1997).

A comparison of the complete mouse and human genome sequences has confirmed the very minor input of small indels to their difference in genome size. Thus, although "the overall loss owing to small indels in ancestral repeats is at least twofold higher in mouse than in human", this contributes only "a small amount (1–2%) to the difference in genome size" (Mouse Genome Sequencing Consortium, 2002). In keeping with this, most pseudogenes in both rodents and humans are now only about 1.2–2.3% shorter than they were when first inserted. It is therefore obvious that "[small] deletions and insertions in murid and human genomes do not contribute significantly to genome size" (Ophir and Graur, 1997). Put another way, a roughly 10% difference in genome size, accumulated over a 75 million year period, has almost nothing to do with small indels. The difference, instead, arises primarily by large deletions in the mouse genome (Mouse Genome Sequencing Consortium, 2002). This is interesting in light of the greater diversity of active transposable elements in the mouse genome versus that of humans.

Similarly, the tiny genome of *Fugu* contains a much higher number of active transposable elements than the human genome (Aparicio et al., 2002). Some authors have suggested that small indels may explain the streamlining of pufferfish genomes (Brainerd et al., 2001), and indeed pufferfishes display deletion rates intermediate between those of *Laupala* and *Drosophila*, in accordance with the sizes of their C-values (Neafsey and Palumbi, 2003). However, the combination of many active transposable elements and a minuscule genome in *Fugu* implies "strong pressure against insertions and for deletions" (Aparicio et al., 2002). Small indel bias may play some role in this group, but it would never be characterized as a "strong pressure" capable of repressing all the active tranposable elements found in the pufferfish genome. In accordance with this, "a decline in the rate of large-scale insertions [rather than increased deletions] is implicated as a probable cause of the genome size reduction in the tetraodontid (smooth) pufferfish lineage" (Neafsey and Palumbi, 2003).

4.4. A closer look at Drosophila

Though generally combined into a single datum, indel bias analyses have actually been performed on two different species of *Drosophila*, namely *D. virilis* and *D. melanogaster*. These species have been separated for about 40 million years—plenty of time for small indel biases to produce the two-fold difference in genome size observed between them. Or so it would be, except that "in all respects, the patterns of spontaneous formation of indels in the two groups are indistinguishable" (Petrov and Hartl, 1998). If anything, DNA loss rates may even be slightly *higher* in the larger-genomed *D. virilis* (Table 1). The very same finding has been reported of smooth (family Tetraodontidae) versus spiny (family Diodontidae) pufferfishes, in which there is no significant difference in deletion rate despite a two-fold range in genome size; indeed, once again the group with the larger genome may actually delete DNA slightly more quickly (Neafsey and Palumbi, 2003). It also bears mentioning in this case that while the spiny pufferfishes do have small genome sizes of about 800 mega (million) base pairs (Mb), the average for all teleosts (excluding known ancient polyploids) is actually only 1000 Mb (Gregory, 2001b). Therefore, while it is true that smooth pufferfishes have the smallest genome sizes among vertebrates (Gregory, 2001b), their DNA loss rates by indel bias do not differ significantly from those of spiny puffers (Neafsey and Palumbi, 2003), whose genome sizes in turn differ little from most other teleosts. In other words, the tiny genomes of smooth puffers probably have nothing to do with high rates of DNA loss, at least as compared to all other bony fishes.

Interestingly, mean intron size is larger in *D. virilis* by about 39%, which Moriyama et al. (1998) point out "is in surprisingly good agreement with the size difference of the two euchromatic genomes [36%]". (*D. virilis* also has longer stretches of microsatellites than *D. melanogaster* (Schlötterer and Harr, 2000), suggesting that the difference in their C-values is the result of variation in several types of non-coding DNA.) However, the use of an average intron size is somewhat misleading, given that while "long" (>80 bp) introns are indeed longer in *D. virilis*, "short" (< 80 bp) introns are actually longer in *D. melanogaster* (Moriyama et al., 1998). In addition, no difference in intron length was detected between *D. virilis* and *D. pseudoobscura*, despite a two-fold difference in C-value; the authors blame this on insufficient sample size (Moriyama et al., 1998).

A similar difference in mean intron size resulting from changes only in long introns has also been observed in birds

Table 1
Data for *relative* deletion rates in the organisms studied so far

Parameter	*Drosophila melanogaster*[a]	*Drosophila virilis*[b]	*Laupala*[c]	Mammals[d]	*Podisma*[e] no hot spots	*Podisma*[e] with hot spots	Rat/mouse[f]	Human[f]	Smooth puffers[g]	Spiny puffers[g]
Genome size (pg)	0.18	0.34	1.93	3.3	16.93	16.93	3.2	3.5	0.4	0.8
Deletions per bp subst.	0.12	0.16	0.07	0.05	0.06	0.28	0.025	0.025	0.04	0.06
Insertions per bp subst.	0.01	0.01	0.02	0.01	0.03	0.04	0.01	0.01	0.03	0.03
Size of deletions (bp)	25	24.3	7	3.2	1.6	1.5	5.91	4.67	19.8	19.1
Size of insertions (bp)	2.8	4	6.5	2.4	1.2	1.1	5.75	8.03	2.7	2.6
Net relative DNA loss (bp per 1 bp subst.)	2.97	3.86	0.36	0.136	0.06	0.376	0.09	0.036	0.7	1.1

The more recent (and much more comprehensive) dataset of Ophir and Graur (1997) is used here, rather than that of Graur et al. (1989) as cited by Petrov (various refs). In addition, rodents and humans are treated separately rather than simply being averaged into a single mammalian datum. Indel data were taken directly from the references listed (note that the *Drosophila* and *Laupala* values have tended to vary slightly among reports), and haploid genome sizes are as given in Gregory (2001b). Net relative DNA loss=(deletion size × deletion ratio) − (insertion size × insertion ratio).

[a] Petrov and Hartl (1998).
[b] Petrov et al. (1996) and Petrov and Hartl (1997).
[c] Petrov et al. (2000).
[d] As presented in various Petrov references based on data taken from Graur et al. (1989).
[e] Bensasson et al. (2001).
[f] Ophir and Graur (1997).
[g] Neafsey and Palumbi (2003) (smooth pufferfishes = family Tetraodontidae, spiny pufferfishes = family Diodontidae).

(Hughes and Hughes, 1995; Gregory, 2002a). While selection for small genome size has been invoked to explain this pattern (e.g., Hughes and Hughes, 1995), it is not immediately obvious why such a difference should be expected under the DNA loss hypothesis. Perhaps this would result because short introns cannot undergo large deletions without affecting adjacent exons. However, this very same limitation on large deletions may also imply that intron lengths in *Drosophila* are influenced primarily by rare but relatively large *insertions* (Moriyama et al., 1998).

Although the difference in intron sizes among *Drosophila* species is often cited in support of the DNA loss hypothesis, this comparison may be inappropriate. For example, analysis of indels in *Drosophila* introns indicates that "the mutational process generates only a modest excess of deletions over insertions" (Comeron and Kreitman, 2000). More importantly, introns do not follow the same indel patterns found in other non-coding sequences (Ptak and Petrov, 2002). According to Ptak and Petrov (2002), "this discrepancy could be explained if deletions, especially long deletions, are more frequently strongly deleterious than insertions and are eliminated disproportionately from intron sequences". However, even the most generous interpretation of their simulation study does not account for the observed pattern. It has also been suggested that intron lengths in *Drosophila* may be maintained by selection rather than indel biases, based on an association between intron size and recombination rate (Carvalho and Clark, 1999; Comeron and Kreitman, 2000). Likewise, in a recent analysis of human intron sequences, Vinogradov (2002) found a correlation between the strength of the indel bias among introns and individual intron sizes, and suggested that the relationship may relate to selection for intron length. On the other hand, a recent study of plants showed that "the rate of indel accumulation in introns was very low, with no evident differences among taxa in this respect" (Wendel et al., 2002b). Evidently, the role of indel bias in shaping intron sizes may vary among taxa, such that its general importance is far from clear.

Finally, and in contrast to the notion that DNA loss rates determine genome size variation among species of *Drosophila*, the difference in C-values between *D. simulans* and *D. melanogaster* has recently been attributed in large part to variation in transposable element copy number, which appears to be influenced by environmental features external to the organism (Vieira et al., 2002). Taken together, these observations would seem to suggest that differences in genome size among closely related species—even those in which DNA loss rates are fast enough to matter—are not shaped primarily by variation in indel patterns.

4.5. Swapping junk for trash?

In summary, there is no single taxonomic level at which the DNA loss mechanism is clearly relevant. For the time being, the important scale to consider remains that of genome size itself. At best, the DNA loss hypothesis can only apply to genomes that delete DNA quickly enough to be relevant in real time, which based on the current dataset seems to apply at values less than 2 pg (i.e., *Drosophila* and pufferfishes, and perhaps *C. elegans*). Above this threshold, variation among taxa is clearly not produced by differences in the rates of small deletions. But even below this level the relevance of DNA loss remains somewhat ambiguous, as exemplified by the case of the *Drosophila* species outlined above.

As discussed elsewhere, traditional mutation pressure theories posit a tendency for genome growth that would continue indefinitely were it not curbed by some countervailing force (see Gregory, 2001a for review). In both the selfish DNA and junk DNA theories, DNA accumulation is

generally believed to be mitigated by the replicational cost incurred by the host organism. Differential tolerances for extraneous DNA among organisms with different replicational requirements would therefore be postulated to account for the variation in genome size among taxa (e.g., Pagel and Johnstone, 1992).

Under the DNA loss hypothesis, differences in the strength of deletional biases may also place a limit on genome expansion in some groups. For example, *Drosophila* is assumed to lack pseudogenes (junk DNA) because it loses them by spontaneous deletion much less slowly than does a cricket or a mammal (e.g., Lozovskaya et al., 1999). In barley, the accumulation of transposable elements (selfish DNA) may be capped by the loss that the sequences themselves engender by the repetitive nature of their terminal sequences. In both of these examples, DNA loss is produced by medium to large deletions. In mammals and orthopterans, spontaneous deletions at the high end of the 400-bp range occur only infrequently or indeed not at all (Graur et al., 1989; Ophir and Graur, 1997; Petrov et al., 2000; Bensasson et al., 2001). Again, DNA loss by deletion bias is functionally nil in both of these groups, meaning only that DNA expansion is not limited by such a mechanism in these animals. Genome size differences within and among the members of these large-genomed groups are therefore not determined by differences in their relative rates of DNA loss by small deletions, although the fact that none has a C-value as small as *Drosophila*'s may indeed be influenced by variations in indel bias.

The primary contribution of the DNA loss hypothesis appears to be that it provides a mechanism in certain organisms for limiting the spread of non-coding DNA that makes no appeals to the replicational costs or phenotypic effects of bulk DNA. This is a useful addition to traditional theories, given that the old view of junk DNA is actually rather implausible (Gregory, 2001a). It is not, however, a new theory of genome size evolution. It is the junk DNA theory revised to include specific mutational mechanisms active only in small-genomed organisms. That is to say, in the absence of deletional containment mechanisms (as opposed to replicational cost limits), DNA will tend to accumulate. Following Sidney Brenner's assertion that "there is an important difference between 'junk DNA' and 'trash DNA': trash DNA is what gets thrown away" (quoted in Petrov and Hartl, 1997), perhaps this updated version should be dubbed the "trash DNA theory" in order to emphasize this key mechanistic distinction.

5. DNA loss in broader perspective

On the basis of the present discussion, it would seem that DNA loss by small indel bias may be relevant under some conditions, but that it is has by no means been established as a dominant force in genome size evolution. In this regard, it is necessary to consider other factors responsible for the modulation of genome sizes, and the way in which DNA loss may interact with these.

5.1. Genome size and the organismal phenotype

It has been recognized since the earliest days of genome size study that nuclear DNA contents are linked in a strong positive way with cell size (reviewed in Gregory, 2001a). This, in turn, seems to provide the basis for relationships between genome size and metabolic rate, developmental rate, body size, and other such parameters, depending on the biology of the group in question (e.g., Gregory et al., 2000; Gregory, 2002a,b, 2003b). Mutation pressure theories typically explain these relationships in coincidental terms. That is, species that develop slowly or have low metabolic rates can simply tolerate the accumulation of more non-coding DNA. The two main optimal DNA theories (nucleotypic and nucleoskeletal) differ in whether they see these relationships as causative or co-evolutionary, but in either case selection on the organism is seen as important in modulating genome sizes via the intermediate of cell size (Cavalier-Smith, 1985; Gregory, 2001a).

Obviously, a strict interpretation of the DNA loss hypothesis would be incompatible with both of these approaches, since it considers genome size to be largely determined by the balance of mutational forces capable of changing genome sizes in either direction. Since most DNA loss proponents rightly accept the validity of these phenotypic correlations (e.g., Hartl, 2000; Petrov, 2001, 2002a), some alternative interpretation must be offered for how genome size can be linked to cellular and organismal parameters.

In general, the DNA loss hypothesis has been taken to imply a "genomes-first" scenario, in which genome size is set by mutational mechanisms, followed by "the rest of the organism adapting to its genome size through co-evolution of other characters and habitat change" (Petrov, 2002a). That changes in genome size can indeed influence evolutionary trajectories in this way is illustrated by intriguing cases such as that of miniaturized plethodontid salamanders. In these animals, there has been a secondary increase in genome size as well as a reduction in body size, with the net effect that they now possess simplified brains made of large, slowly dividing neurons packed within a tiny brain case. Rather than simply reducing genome size to prevent this effect on brain complexity, these salamanders have changed from a computationally demanding active predation strategy to a lie-in-wait strategy, including the evolution of a specialized projectile tongue (e.g., Roth et al., 1997).

Of course, the genomes of these salamanders are far larger than even that of *Podisma* grasshoppers, so factors other than DNA loss rate must be invoked to explain this increase in DNA content. In dipterans versus orthopterans, on the other hand, genome size differences may be influenced by faster DNA loss in the former. This may be of particular relevance, since the major defining feature of genome size variation in insects appears to be the presence or absence of complete

metamorphosis during the developmental program (Gregory, 2002a). Holometabolous insects, such as flies, have genomes smaller than 2 pg, whereas many hemimetabolous groups, most notably grasshoppers, may exceed 2 pg by a wide margin. It could be that an early ancestor to the holometabolous orders had a high rate of DNA loss which kept its genome size small, thereby allowing it the option of undergoing metamorphosis. Hemimetabolous orders, with their larger genome sizes, may have been unable to evolve complete metamorphosis because of the developmental effects of large quantities of DNA. Of course, natural selection for smaller genomes in response to developmental constraints is an (at least) equally likely explanation, but the possibility that DNA loss rates helped to shape such a major evolutionary divergence is very interesting.

Similar genomes-first processes have been proposed for plants as well: "Although microalterations in genome size may not provide sufficient material for significant degrees of natural selection, plants must deal with the genome contents that have been generated by mechanisms of shrinkage and expansion. Vastly different genome sizes, possibly arrived at without selection during their incremental progression, will influence how a plant prospers in any given environment" (Bennetzen, 2002). For example, Petrov (2002a) raises the observation that perennials have larger genomes than annuals, and explains this in terms of an annual plant necessarily becoming perennial as its genome grows slowly in size in the absence of sufficient DNA loss.

But the important question is: Does the relationship *always* proceed in this way, with genome size set first and the organism simply adjusting to it? Taking Petrov's (2002a) example of annual versus perennial plants, it would seem that the answer is no. In fact, the shift in developmental lifestyle in plants is usually in the opposite direction, with perennials becoming annuals as they move to new habitats with harsh environmental conditions (see Gregory, 2002a and references therein). Importantly, there is now phylogenetic evidence that genome size reduction occurs as part of this process (e.g., Watanabe et al., 1999). What is not clear, however, is the direction of causation—i.e., does an initial chance reduction in genome size allow the shift to annual lifestyle, or does the organismal change prompt an active genome size reduction? Were genome size shrinkage not so incredibly slow in organisms with genomes of the size found in perennial plants, it might have been that a change in indel spectra could have instigated the shift to an annual lifestyle. However, as it stands this is extremely unlikely. On the other hand, the nucleotypic interpretation favoured by most botanists, that genome size must be actively reduced to allow the adoption of an annual lifestyle, remains entirely plausible (e.g., Bennett, 1987; Watanabe et al., 1999).

Other examples are far less ambiguous. For example, there is reason to believe that flighted birds and rapidly metamorphosing amphibians (e.g., frogs inhabiting ephemeral pools) must maintain small genomes (Hughes, 1999; Gregory, 2002a,b). Traditional explanations would suggest selection on the organism level to cap the spread of transposable elements and the accumulation of pseudogenes, and indeed there is evidence to this effect (e.g., Baker et al., 1992; Hughes, 1999). From the DNA loss perspective, the causation would proceed in the reverse direction, with genome size set as small or large on the basis of arbitrarily determined loss patterns, and the organism left to adapt to the large or small genome. However, it is notable that in both flightless birds and neotenic amphibians genome sizes are seen to increase along with the shift in lifestyle (e.g., Martin and Gordon, 1995; Hughes, 1999). In birds, there is some evidence that genomes were small prior to the evolution of flight (Waltari and Edwards, 2002), which could be compatible with a mutational genomes-first model. However, the significant difference in genome size between strong fliers and flightless birds is too small to be interpreted in terms of a change in DNA loss rates causing genomes to grow and the birds to lose flight, and instead they support the more parsimonious notion of relaxed selective constraints in secondarily flightless taxa (Hughes, 1999). In amphibians, fossil cell size data make it clear that an expansion in genome size followed the simplification of development (Thomson and Muraszko, 1978), and not the reverse as necessarily assumed by a strict DNA loss model.

5.2. Other mechanisms of genome size change

Deletion bias, even in the fastest cases, is still a very slow process. Because each insertion of a transposon adds an entire element, whereas deletion bias only whittles these away gradually, large insertions must be few and far between if DNA loss is to have any long-term effects (see also Neafsey and Palumbi, 2003). Since transposable elements are generally not nearly so subdued, differences in the rates of insertion (or, more precisely, fixation of insertions) among species are probably much more important in determining genome size diversity than variation in small deletion rates. This is very clearly true of maize (SanMiguel et al., 1998) and apparently also of humans versus chimpanzees (Liu et al., 2003) and smooth versus spiny pufferfishes (Neafsey and Palumbi, 2003), and may just as easily apply to even the fastest DNA losers like *Drosophila* (Vieira et al., 2002).

The DNA loss hypothesis deals with a specific mechanism of genome size reduction—namely, biases towards small deletions—and a rejection of this approach does not imply a belief that genome sizes can only increase through time. Other mechanisms of DNA loss, including large deletions, unequal crossing over, elimination of entire chromosomes or large segments thereof (e.g., following polyploidy), or active streamlining in response to selection, can all produce significant reductions in DNA content. Nevertheless, the most commonly accepted mechanisms of genome size change do involve increases in DNA content. In some cases, these forces can be a great deal more powerful than any proposed mechanisms of DNA loss. Thus, polyploidization, large-scale duplications, and surges of trans-

posable element activity can all increase DNA content very rapidly. Again, according to the mutational equilibrium model, these processes represent noise around a long-term stable value. However, even with a relatively small genome size like that of maize, a single rapid increase in genome size by transposable element proliferation is likely to be permanent unless more powerful reductive forces come into play, considering how slowly the small indel bias mechanism operates in real time.

6. Alternative interpretations

It is simply a truism that the observed genome size is the result of a balance between the rates of DNA gain and loss, so the identification of a new DNA loss mechanism does not, in itself, resolve the enigma of genome size evolution. Even if one accepts a fully mutational model of genome size, the crucial question will remain: *Why* do some organisms lose (or gain) DNA quickly while others do not?

6.1. Rapid replication and sloppy slippage

In *Drosophila*, 1-bp deletions—what could be called "very small errors"—account for 26% of the observed deletions (Petrov and Hartl, 1998). By contrast, very small errors make up 50% of rodent pseudogene deletions, and in humans the figure is 57% (Graur et al., 1989). As discussed below, the differences among insects in average deletion rate are, in reality, differences in the frequency of deletions greater than 15 bp, or what could be considered "larger errors". Indeed, *Podisma* shows no evidence of deletions larger than 4 bp (Bensasson et al., 2001). In one manner of speaking, the only observation to be explained in the entire DNA loss dataset is why *Drosophila* (and perhaps *Caenorhabditis* and *Fugu*) exhibits more "larger errors" (e.g., due to such processes as replication slippage) than the other species.

It has long been asserted that animals with short generation times accumulate mutations more quickly than those with longer life cycles (e.g., Wu and Li, 1985). More generations per unit time means more opportunities for errors per unit time and thus a speedier rate of mutation *accumulation*. But what if faster generation/replication time also *caused* more frequent and more severe errors? Developmental rate is known to correlate inversely with genome size in many organisms, including some arthropods (Gregory, 2002a). As Petrov has noted, "organisms like fruit flies may be careless about copying junk DNA when replicating their chromosomes, giving them compact, junk-free genomes. Others, like onions, may faithfully reproduce everything, resulting in a cluttered and junky genome" (quoted in Cromie, 2000). Insofar as this replicational carelessness in *Drosophila* could be related to its extraordinarily rapid life cycle (~ 2 weeks), it seems that this factor needs to be controlled before it can be dismissed as an explanation for the observed correlation between deletion rate and genome size. This could be accomplished by examining indel biases in an insect such as the common streamdwelling water strider, *Gerris remigis* (Hemiptera: Gerridae). While this critter has a significantly longer generation time than *Drosophila* (development in about 2 months, often only one generation per year with the adults overwintering at northern latitudes; Cheng and Fernando, 1970), it has a genome size slightly smaller than that of *D. virilis* (~ 0.3 pg; Gregory, unpublished data). The reverse control, an insect with a large genome but a rapid life cycle, will probably be rather hard to find (Gregory, 2002a).

6.2. Unequal patterns of unequal crossing-over

It might be argued that replication slippage does not produce deletions on the order of several hundred base pairs, and is therefore not responsible for the DNA loss patterns found in *Drosophila*. In fact, it seems likely that "deletions are generated by more than one mechanism: one mechanism is relatively constant between species and generates predominantly small (single-nucleotide) deletions (e.g., as expected through slipped-strand mutation), while a second mechanism generates a broader range of deletion sizes (e.g., as expected through unequal sister chromatid exchange or unequal crossing-over) and is more active in organisms with smaller genomes" (Bensasson et al., 2001). If so, then sloppy replication in insects with rapid life cycles may not provide a satisfactory explanation for their higher rates of DNA loss.

In comparing the distributions of deletion sizes in *Drosophila* versus mammals, it becomes clear that "the only variable parameter ... is the size distribution of deletions and, in particular, presence or absence of deletions longer than 5 bp" (Petrov, 2002b), with *Drosophila* deleting fragments substantially larger than this, but mammals not. If the limit for deletions produced by replication slippage is about 5 bp, then this discontinuity would most likely reflect the transition from one deletional mechanism to another. Thus, the question to be addressed would become even more specific: Why does unequal crossing over cause larger deletions in *Drosophila* but not in the other animal species?

In plants, the absolute rate of recombination increases with chromosome length, and therefore also with genome size in most cases. However, the *relative* rate of recombination (i.e., chiasma frequency per unit DNA) decreases with increasing genome size. Specifically, relative recombination rate is much higher in organisms with very small genomes, but this effect drops off significantly at C-values larger than 3 pg (Rees and Durrant, 1986). Intriguingly, many large deletions, probably produced by unequal crossing-over, occur in *Drosophila* (0.18 pg), a few in cricket (1.9 pg), and none in animals with genomes larger than mammals (>3 pg). If the relative DNA loss rate (i.e., the likelihood of experiencing large deletions) is linked to the relative frequency of recombination, which in turn is dependent on genome size, then this could also provide a partial alternative explanation for the DNA loss rate correlation.

6.3. Differences among sequences examined

It is worth considering the potential problems involved in choosing to analyze different types of sequences in the various species so far studied. Since the rates of replication slippage and recombination are both dependent on the presence of repetitive sequences, differences in repetition within pseudogene/transposon/numt sequences could directly impact the overall estimated rates of loss. For example, in *Podisma* there is a "mononucleotide repeat within the 643-bp pseudogene sequence studied that acts as a strong hot spot for insertions and deletions" (Bensasson et al., 2001). Citing the lack of such hot spots in the previously studied species, Bensasson et al. (2001) ignore it in their analysis. However, when data for this hot spot are included, *Podisma* actually has a calculated deletion rate *higher* than that of *Laupala*, despite a 10-fold larger genome (Table 1). Indeed, there are "major differences among the taxa studied in the sizes and strengths of hot spots [which] appear to have arisen as a result of the differences in the lengths and compositions of the original sequence studied" (Bensasson et al., 2001). And again, there is a still-unexplained difference in indel dynamics between transposable elements and pseudogenes on the one hand and introns on the other (Ptak and Petrov, 2002).

Perhaps most importantly, the recent analysis of Blumenstiel et al. (2002) has shown that while the deletion-to-insertion ratio in the *Drosophila* genome at large is about 3.6, the ratio in *Helena* elements (the sequence upon which the bulk of the DNA loss hypothesis is based) it is at least 7.1, and perhaps as high as 9.1. In plain terms, "*Helena* may contain regions that are more deletion-prone" (Blumenstiel et al., 2002). Moreover, the inclusion of additional sequences revealed that insertions were considerably more common in *Drosophila* than originally estimated based only on *Helena* elements. With the inclusion of this expanded dataset, the deletion bias in *Drosophila* was found to be only half as strong as initially reported (Blumenstiel et al., 2002). Likewise, a roughly two-fold difference was found in pufferfishes depending on whether pseudogenes or dead-on-arrival transposable elements were assayed (Dasilva et al., 2002; Neafsey and Palumbi, 2003).

While the claim is not being advanced here that the entire DNA loss relationship is an artifact of fortuitous sequence choice, it is clear that this issue warrants more consideration than it has been given thus far.

6.4. Is Drosophila just a freak?

Using the more recent rodent and human data given in Ophir and Graur (1997) analyzed separately (see below), and omitting the extreme case of *Drosophila*, the relationship between genome size and relative deletion rate is only marginally significant even when log-transformed (Fig. 1).

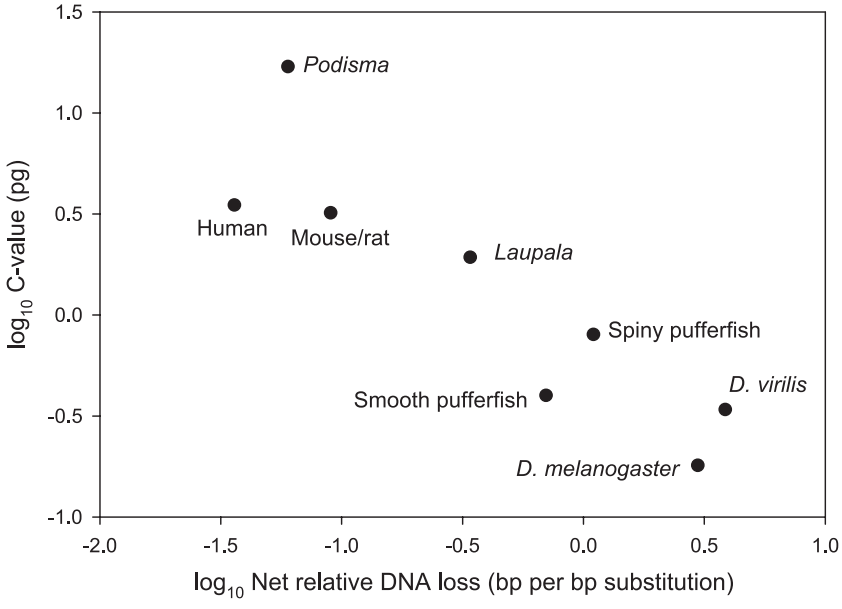

Fig. 1. The relationship between genome size and relative DNA loss rate, using the updated dataset given in Table 1 ($r^2 = 0.81$, $p < 0.0025$). When the extreme point from *Drosophila* is removed, the relationship is only marginally significant even when log-transformed ($r^2 = 0.68$, $p = 0.045$). When the highly skewed deletion size data are log-transformed prior to calculating the average (i.e., rather than using simple arithmetic means), the relationship is much weaker (see Table 2). Note that the two-fold differences in genome size between *D. melanogaster* and *D. virilis* and between smooth (family Tetraodontidae) and spiny (family Diodontidae) pufferfishes do not correspond to differences in DNA loss rate. Note also that the *Drosophila* and pufferfish values may vary up to two-fold according to which type of sequence is analyzed (Blumenstie, 2002; Neafsey and Palumbi, 2003). Moreover, when the updated mammalian dataset of Ophir and Graur (1997) is used instead of that of Graur et al. (1989), it is apparent that humans delete DNA more slowly than *Podisma* grasshoppers, despite possessing a five-fold smaller genome. Taken together, these findings raise considerable doubts about the strength of any correlation between DNA loss rate by small indel bias and genome size.

The remaining dataset still includes a genome size range of nearly 10-fold, and it is notable that relationships with cell size and metabolic rate (which are far noisier than the proposed causative relationship with DNA loss rate) can still be identified in both mammals and birds over only a four- and two-fold range, respectively (Vinogradov, 1995; Gregory, 2002a,b). Perhaps the addition of more data will rescue the fruitfly-free correlation, but until then it is important to note that there is no strong relationship without *Drosophila*. This supports the notion that DNA loss rate can, at best, explain the small genome of *Drosophila* but not the very different genome sizes of the other taxa. It also raises the question: could there just be something peculiar about *Drosophila*'s genome? Is there reason to believe it is unusually error prone?

Indeed there is. A recent comparison between distantly related *Drosophila* species revealed no less than 114 fixed paracentric inversions, which amounts to roughly one fixed inversion every million years (Ranz et al., 2001). This rate of chromosomal disruption is one of the highest so far found in any eukaryote, being at least five times faster than even the most dynamic plant genomes (specifically, the *Arabidopsis–Brassica* clade), and higher than the rate in mammals by two orders of magnitude (Ranz et al., 2001; González et al., 2002). As such, the "extraordinarily malleable" genomes of *Drosophila* make it a clear outlier in the current DNA loss rate analysis. Importantly, if *C. elegans* proves to fit on the DNA loss rate-genome size curve once reliable data are collected for this species, this alternative interpretation may actually gain support, since its genome is the only one known to be even more erratic than that of *Drosophila* (Coghlan and Wolfe, 2002).

6.5. Selection for small genome size

Proponents of the DNA loss hypothesis have gone to great lengths to dismiss the rather extreme view that each individual indel is under selection in *Drosophila* (Petrov et al., 1996; Petrov and Hartl, 1997, 2000; Moriyama et al., 1998; Lozovskaya et al., 1999; Hartl, 2000; Petrov, 2001, 2002b). The basic argument is as follows. If selection for small genome size directly influences indel patterns, then there ought to be a correlation between the age of a "pseudogene" (i.e., number of substitutions) and the average size of deletions found in it. This is because large deletions contribute more to shrinking the genome and therefore should be favoured by selection. Yet no such correlation exists among species. (There does, however, appear to be a relationship between indel bias strength and segment age among human introns; Vinogradov, 2002.) Although some objections have been raised to this approach (e.g., Charlesworth, 1996), it seems fully reasonable to accept the conclusion that selection does not differentiate among individual small indels.

Importantly, there are other ways to invoke selection for small genome size that are not so naïve. For example, it is easy to imagine a genome-wide tendency to indiscriminately delete non-coding DNA that is forged by selection for a small genome. Such a mechanism may in fact be operating in bacteria (and, perhaps, in *Drosophila*). This broader selectionist view has not been challenged by DNA loss supporters:

> We wish to distinguish between two versions of the small-genome hypothesis. One is that the number and size of deletions in our data are due to [direct] effects of selection. This version we reject because of the evidence [that large deletions are not more frequent than small ones, as would be expected since the latter would contribute more to genome size]. The more general version of the hypothesis is that selection for a smaller genome size is ultimately the reason why the *Drosophila* genome has such a high rate of deletion. The targets of such selection could be enzymes involved in DNA replication, repair, or recombination. Our data do not bear on this latter version of the small-genome hypothesis, but to us it seems plausible. (Petrov and Hartl, 1997)

Optimal DNA theorists will undoubtedly agree on all counts. Obviously, if this turns out to be the case then the correlation between deletion rate and genome size will have resulted from a generalized selection for small genome size. Moreover, this means that a high rate of DNA loss by deletion bias need not even be the mechanism responsible for producing a small genome size, but could be merely a byproduct of the processes that are.

In their recent analysis, Blumenstiel et al. (2002) showed that *Drosophila* transposable elements are more concentrated in heterochromatic regions of low recombination, but that deletion rates are identical in both euchromatic and heterochromatic regions of the genome. As part of this study, these authors reaffirmed the unlikely role of selection in moderating each tiny deletion event, and extended this to include selective models relating to ectopic recombination. However, they noted that "although we are able to reject the suggestion that selection is acting strongly on the predominant small deletions, there is some evidence that deletions larger than 400 bp may, in fact, be advantageous". Granted that the DNA loss hypothesis currently focuses on deletions smaller than 400 bp, this is not a direct argument for selection in explaining the patterns of DNA loss upon which it is based. However, 400 bp is an arbitrary cut-off, and it could be that most DNA loss in *Drosophila* actually occurs by deletions larger than this. This would seem to be a reasonable expectation, since even below 400 bp, the presence of deletions at the higher end of the spectrum is the only thing that produces a faster "rate" of DNA loss in *Drosophila* versus the other animals. Selection may indeed fail to discriminate among very small deletions, but for the most part these are not relevant to differences in genome size in any case.

7. Some cautions regarding the current DNA loss dataset

All of the above discussions have been presented on the assumption that the current dataset accurately reflects the relationship between relative DNA loss rate and genome size. However, there are reasons to question this reported correlation, which will be discussed in the following sections.

7.1. Of mice and men

Because Graur et al. (1989) did not list rates or sizes of indels in rodents and humans, Petrov et al. (1996) were forced to calculate them from the raw data for comparison with their estimates from *Drosophila*. In so doing, they combined the data from rats/mice and humans to give a single "mammal" value, even though Graur et al. (1989) indicated that deletions occur significantly more often in rodents than in humans. As mentioned previously, this average mammalian deletion rate was slower than *Drosophila* and cricket, but faster than grasshopper.

These are not the only (or even the best) mammalian data now available. In fact, shortly after the initial *Drosophila* study, Ophir and Graur (1997) presented a greatly expanded survey of mammalian processed pseudogenes (93 from human, 63 from rats and mice—roughly triple the amount of data presented in the original study). Yet, the data from the older study continue to be cited and compared with the insect values. This failure to update the mammalian data is significant because the sizes of mammalian deletions and (especially) insertions, as well as the relative insertion rates, all now appear to have been underestimated in the dataset consistently used by DNA loss proponents. In fact, using the updated values given in Ophir and Graur (1997), it seems that humans delete DNA slightly more *slowly* than *Podisma*, even though the latter has five times more DNA in its genome (Table 1).

7.2. Some statistical considerations

The current DNA loss dataset is based on the calculation of averages. Relative DNA loss rates are calculated using mean values for indel sizes and frequencies, and then compared against average substitution rates. Rodent and human indel patterns are averaged, as are those of *Drosophila melanogaster* and *D. virilis*. It seems important to consider what effect all this averaging has on the resulting relationship.

To reiterate, there is very little difference in the actual *rates* of DNA deletion among the animals studied thus far; the observed differences result from variation in the average *sizes* of deletions. Specifically, "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*. In particular, the rates of deletions of 1–5 bp are indistinguishable in *Drosophila* and mammals" (Petrov and Hartl, 2000). The same goes for crickets: "Most of the difference in DNA loss [in *Drosophila* vs. *Laupala*] is due to *Laupala* having a much smaller fraction of deletions larger than 15 bp. For deletions smaller than 15 bp, the rates of deletions per substitution are indistinguishable" (Petrov et al., 2000).

Even a single very large deletion in one species could lead to a much larger *average* deletion size and therefore a much faster estimated rate of loss. In fact, this is not far from the actual situation. Based on an examination of the frequency distributions provided by Bensasson et al. (2001) for deletion sizes in the three insects studied, it is apparent that the entire difference in DNA loss "rate" between crickets and grasshoppers is a product of only seven mid-sized deletions which occurred in *Laupala* but not in *Podisma*.

More importantly, the *Drosophila* and *Laupala* distributions are obviously highly skewed—with standard deviations nearly twice as large as the means—which raises doubts regarding the validity of using simple arithmetic averages to calculate deletion rates. Had the deletion data been log-transformed prior to calculating the average (rather than the reverse), the resulting estimated deletion sizes would have been less biased by the few very large deletions at the tail of the highly skewed distribution (Gregory, 2003a). Indeed, geometric means calculated using $\log(x+1)$-transformed deletion size data are much lower for *Drosophila* and *Laupala* (Table 2), which would make the resulting pseudogene half-lives considerably longer than estimated above. More seriously, the calculation of geometric means brings the *Laupala* and *Podisma* rates very close together, despite their nearly 10-fold difference in genome size (Table 2). (Note that these calculations were done assuming the same average insertion sizes, the raw data for which were unavailable for log-

Table 2
Effects of $\log(x+1)$ transformation on mean deletion sizes and subsequent loss rate calculations in insects

Parameter	*Drosophila*	*Laupala*	*Podisma*
Ave deletion size, arithmetic mean (bp)	35.4	6.7	1.6
Standard deviation of arithmetic mean	64.7	14.6	1.1
Ave deletion size, geometric mean (bp)	11.2	3.1	1.4
Net relative loss, arithmetic mean (bp/bp subst)	4.6	0.34	0.07
Net relative loss, geometric mean (bp/bp subst)	1.4	0.08	0.06

In both *Drosophila* and *Laupala*, deletion size data distributions are highly skewed and have standard deviations roughly twice as large as the means, indicating that the current use of arithmetic means to calculate net loss rate is questionable. Both arithmetic and geometric means were calculated using the raw deletion size data given in Fig. 2 of Bensasson et al. (2001). Net relative loss rates were calculated using the additional data and equations listed in Table 1.

transformation. However, since large insertions are far less common than large deletions within the relevant size range, this may not have any substantial effect on the calculations presented here.)

8. Concluding remarks

8.1. The future of DNA loss

Notwithstanding any doubts about the accuracy of the calculated loss rates (Tables 1 and 2) and generality of the correlation within the existing dataset (Fig. 1), it is intriguing that there may be an identifiable relationship between genome size and DNA loss rate across two orders of magnitude. However, even taking the current dataset at face value, there are several crucial questions that must be answered empirically before the DNA loss hypothesis can be applied generally to the question of genome size. These include: (1) How much variation in genome size is there among species with the same DNA loss rate? (2) How much variation in DNA loss rate is there among species with the same genome size? (3) Can DNA loss rate data alone accurately predict unknown genome sizes? It should, if it is a major determinant. (4) On what temporal and taxonomic scales is the indel bias effect presumed to be important? (5) Do different sequences examined from the same species give different estimates of DNA loss rate? It should be clear that until the controls needed to answer these questions have been implemented, the DNA loss dataset can be considered only very preliminary. At present, the data are far too limited to justify any assumptions about the generality of the phenomenon, and for this reason authors should be especially cautious when applying the concept of DNA loss to groups for which no data yet exist.

8.2. Indel bias and the C-value enigma

Even if the data underlying the DNA loss hypothesis are greatly expanded, properly controlled for extraneous variables, calculated with the best available data using appropriate statistical methods, and still found to correlate strongly with genome size, this would not necessarily mean that indel bias provides a solution to the problem of genome size variation. Of course, a statistically significant correlation does not necessarily indicate a causative, or even a biologically relevant, relationship. It is also clear, even based on the small dataset currently available, that the DNA loss hypothesis has only limited input into the generation of large-scale variation in genome size among eukaryotes. In particular, it seems to be of only minor significance to the really interesting questions of the C-value enigma. For example, it may help to explain why *Drosophila* has a smaller genome than a cricket (though, as discussed above, there are additional developmental factors to consider; see also Gregory, 2002a). It cannot, however, explain why the human genome is larger than a cricket's, or why a grasshopper's is larger than a human's. To be sure, the fact that mountain grasshoppers have a genome size five times larger than humans is a very nice illustration of the type of thing that must be explained in order to solve the C-value enigma. But because the input of the DNA loss mechanism in real time is essentially zero in both groups, this difference must be caused entirely by other factors (probably *large* insertions and deletions). Selection is also of potential relevance to this question, given the obvious suggestion that humans, as homeothermic vertebrates, could not tolerate such a large genome on the grounds of cell size/metabolic rate effects (see Vinogradov, 1995; Gregory, 2000, 2002b). Indeed, it is not particularly surprising that some organisms have very small genomes—in fact, this is the null hypothesis. The interesting question is why some species have very large genomes, and on this issue the DNA loss hypothesis apparently has very little to say—except, perhaps, as a new twist on the traditional junk DNA theory.

With regards to the mutational equilibrium model, it must be concluded that even if the current DNA loss data are accepted as they are despite some important concerns, these are too few to justify discussions of scaling functions as evidence for or against any particular theory of genome size change. Moreover, there are alternative interpretations to the current dataset that are compatible with previous theories, suggesting that the observed data do not require the development of a totally new model. Finally, the model itself requires some assumptions that, while certainly testable, appear rather implausible based on the current understanding of genome evolution.

8.3. The power of pluralism

While the long-outdated term "C-value paradox" implies a one-dimensional problem, the "C-value enigma" has been explicitly formulated as a series of independent puzzles (Gregory, 2001a). This distinction is important, because the framework for the problem determines the types of solutions sought. When viewed as a complex puzzle (i.e., an "enigma"), it is clear that a complete model of genome size evolution will require a combination of several explanatory approaches. This includes, but is clearly not limited to, mechanistic explanations dealing with the ways in which DNA is gained and lost. Similarly, it is important to recognize and explain the relationships between genome size and cellular and organismal features. However, neither one of these approaches, by itself, provides a solution to the C-value enigma. The DNA loss hypothesis, in particular, should not be viewed as "the" solution to the overall enigma. What is needed is a pluralistic outlook open to insights derived from many different biological disciplines, from molecular biology to ecology. Such an intriguing and long-standing biological riddle demands nothing less.

Acknowledgements

Supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) post-doctoral fellowship and the NSERC Howard Alper Post-Doctoral Prize. I thank S. Adamowicz, T. Crease, D. Graur, S. Johnston, H. Robertson, A. Schulman, A. van Wijnen, and the anonymous reviewers for providing helpful comments and criticisms on an earlier draft of the paper, and D. Petrov for stimulating discussions.

References

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., et al., 2002. Whole genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297, 1301–1310.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Arkhipova, I., Meselson, M., 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. U. S. A. 97, 14473–14477.

Baker, R.J., Maltbie, M., Owen, J.G., Hamilton, M.J., Bradley, R.D., 1992. Reduced number of ribosomal sites in bats: evidence for a mechanism to contain genome size. J. Mammal. 73, 847–858.

Beaton, M.J., Cavalier-Smith, T., 1999. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. Proc. R. Soc. Lond., B 266, 2053–2059.

Bennett, M.D., 1987. Variation in genomic form in plants and its ecological implications. New Phytol. 106 (Suppl.), 177–200.

Bennett, M.D., Leitch, I.J., Hanson, L., 1998. DNA amounts in two samples of angiosperm weeds. Ann. Bot. 82 (Suppl. A), 121–134.

Bennetzen, J.L., 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115, 29–36.

Bennetzen, J.L., Kellogg, E.A., 1997. Do plants have a one-way ticket to genomic obesity? Plant Cell 9, 1509–1514.

Bensasson, D., Petrov, D.A., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. 18, 246–253.

Betrán, E., Long, M., 2002. Expansion of genome coding region by acquisition of new genes. Genetica 115, 65–80.

Blanc, G., Hokamp, K., Wolfe, K.H., 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. 13, 137–144.

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

Brainerd, E.L., Slutz, S.S., Hall, E.K., Phillis, R.W., 2001. Patterns of genome size variation in tetraodontiform fishes. Evolution 55, 2363–2368.

Bullock, D., Rayburn, A., 1991. Genome size variation in southwestern US Indian maize populations may be a function of effective growing season. Maydica 36, 247–250.

Carvalho, A.B., Clark, A.G., 1999. Intron size and natural selection. Nature 401, 344.

Cavalcanti, A.R.O., Ferreira, R., Gu, Z., Li, W.-H., 2003. Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. J. Mol. Evol. 56, 28–37.

Cavalier-Smith, T., 1985. Cell volume and the evolution of eukaryotic genome size. In: Cavalier-Smith, T. (Ed.), The Evolution of Genome Size. John Wiley & Sons, Chichester, pp. 104–184.

Cavalier-Smith, T., 2002. Nucleomorphs: enslaved algal nuclei. Curr. Opin. Microbiol. 5, 612–619.

Cavalier-Smith, T., Beaton, M.J., 1999. The skeletal function of non-genic nuclear DNA: new evidence from ancient cell chimaeras. Genetica 106, 3–13.

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

Cheng, L., Fernando, C.H., 1970. The Water-Striders of Ontario (Heteroptera: Gerridae). Royal Ontario Museum Life Sci. Misc. Pubs, Toronto, Ontario.

Claverie, J.-M., 2000. What if there are only 30,000 human genes? Science 291, 1255–1257.

Coghlan, A., Wolfe, K.H., 2002. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res. 16, 857–867.

Comeron, J.M., 2001. What controls the length of noncoding DNA? Curr. Opin. Genet. Dev. 11, 652–659.

Comeron, J.M., Kreitman, M., 2000. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics 156, 1175–1190.

Cromie, W.J., 2000. Why onions have more DNA than you do. Harvard Univ. Gazette Feb. 10.

Dasilva, C., Hadji, H., Ozouf-Costaz, C., Nicaud, S., Laillon, O., Weissenbach, J., Roest Crollius, H., 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. Proc. Natl. Acad. Sci. U. S. A. 99, 13636–13641.

de Jong, W.W., Rydén, L., 1981. Causes of more frequent deletions than insertions in mutations and protein evolution. Nature 290, 157–159.

Devos, K.M., Brown, J.K.M., Bennetzen, J.L., 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 12, 1075–1079.

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

Frank, A.C., Amiri, H., Andersson, S.G.E., 2002. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica 115, 1–12.

Friedman, R., Hughes, A.L., 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. 11, 373–3871.

García-Martínez, J., Martínez-Izquierdo, J., 2003. Study on the evolution of the *grande* retrotransposon in the *Zea* genus. Mol. Biol. Evol. 20, 831–841.

Gaut, B.S., 2002. Evolutionary dynamics of grass genomes. New Phytol. 154, 15–28.

Gilson, P.R., 2001. Nucleomorph genomes: much ado about practically nothing. Genome Biol. 2, 1002.1–1002.5.

Gilson, P.R., McFadden, G.I., 2002. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. Genetica 115, 13–28.

González, J., Ranz, J.M., Ruiz, A., 2002. Chromosomal elements evolve at different rates in the *Drosophila* genome. Genetics 161, 1137–1154.

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

Gregory, T.R., 2000. Nucleotypic effects without nuclei: genome size and erythrocyte size in mammals. Genome 43, 895–901.

Gregory, T.R., 2001a. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. 76, 65–101.

Gregory, T.R., 2001. Animal Genome Size Database. http://www.genomesize.com.

Gregory, T.R., 2002a. Genome size and developmental complexity. Genetica 115, 131–146.

Gregory, T.R., 2002b. A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class Aves. Evolution 56, 121–130.

Gregory, T.R., 2003a. Is small indel bias a determinant of genome size? Trends Genet. 19, 485–488.

Gregory, T.R., 2003b. Variation across amphibian species in the size of the nuclear genome supports a pluralistic, hierarchical approach to the C-value enigma. Biol. J. Linn. Soc. 79, 329–339.

Gregory, T.R., Hebert, P.D.N., Kolasa, J., 2000. Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. Heredity 84, 201–208.

Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P., Gerstein, M., 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucleic Acids Res. 31, 1033–1037.

Hartl, D.L., 2000. Molecular melodies in high and low *C*. Nat. Rev., Genet. 1, 145–159.

Hedges, S.B., Kumar, S., 2002. Vertebrate genomes compared. Science 297, 1283–1285.

Hughes, A.L., 1999. Adaptive Evolution of Genes and Genomes. Oxford Univ. Press, Oxford, UK.

Hughes, A.L., Hughes, M.K., 1995. Small genomes for better flyers. Nature 377, 391.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

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

Jockusch, E.L., 1997. An evolutionary correlate of genome size change in plethodontid salamanders. Proc. R. Soc. Lond., B 264, 597–604.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. U. S. A. 97, 6603–6607.

Kirik, A., Salomon, S., Puchta, H., 2000. Species-specific double-strand break repair and genome evolution in plants. EMBO J. 19, 5562–5566.

Leitch, I.J., Chase, M.W., Bennett, M.D., 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. Ann. Bot. 82 (Suppl. A), 85–94.

Liò, P., 2002. Investigating the relationship between genome structure, composition, and ecology in prokaryotes. Mol. Biol. Evol. 19, 789–800.

Liu, G., NISC Comparative Sequencing Program, Zhao, S., Bailey, J.A., Sahinalp, S.C., Alkan, C., Tuzun, E., Green, E.D., Eichler, E.E., 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res. 13, 358–368.

Lozovskaya, E.R., Nurminsky, D.I., Petrov, D.A., Hartl, D.L., 1999. Genome size as a mutation selection-drift process. Genes & Genet. Syst. 74, 201–207.

Mark Welch, D.B., Meselson, M., 1998. Measurements of the genome size of the monogonont rotifer *Brachionus plicatilis* and of the bdelloid rotifers *Philodina roseola* and *Habrotrocha constricta*. Hydrobiologia 387, 395–402.

Mark Welch, D.B., Meselson, M., 2003. Oocyte nuclear DNA content and GC proportion in rotifers of the anciently asexual Class Bdelloidea. Biol. J. Linn. Soc. 79, 85–91.

Martin, C.C., Gordon, R., 1995. Differentiation trees, a junk DNA molecular clock, and the evolution of neoteny in salamanders. J. Evol. Biol. 8, 339–354.

Mira, A., Ochman, H., Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17, 589–596.

Moriyama, E.N., Petrov, D.A., Hartl, D.L., 1998. Genome size and intron size in *Drosophila*. Mol. Biol. Evol. 15, 770–773.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Neafsey, D.E., Palumbi, S.R., 2003. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome Res. 13, 821–830.

Ohno, S., 1972. So much "junk" DNA in our genome. In: Smith, H.H. (Ed.), Evolution of Genetic Systems. Gordon and Breach, New York, pp. 366–370.

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

Orel, N., Puchta, H., 2003. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. Plant Mol. Biol. 51, 523–531.

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

Pagel, M., Johnstone, R.A., 1992. Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. Proc. R. Soc. Lond., B 249, 119–124.

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

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

Petrov, D.A., 2002a. Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. 61, 533–546.

Petrov, D.A., 2002b. DNA loss and evolution of genome size in *Drosophila*. Genetica 115, 81–91.

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

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

Petrov, D.A., Hartl, D.L., 2000. Pseudogene evolution and natural selection for a compact genome. J. Heredity 91, 221–227.

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

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., Shaw, K.L., 2000. Evidence for DNA loss as a determinant of genome size. Science 287, 1060–1062.

Poggio, L., Rosato, M., Chiavarino, A.M., Naranjo, C.A., 1998. Genome size and environmental correlations in maize (*Zea mays* ssp. *mays*, Poaceae). Ann. Bot. 82 (Suppl. A), 107–115.

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

Rabinowicz, P.D., 2000. Are obese plant genomes on a diet? Genome Res. 10, 893–894.

Ranz, J.M., Casals, F., Ruiz, A., 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. Genome Res. 11, 230–239.

Rayburn, A.L., Auger, J.A., 1990. Genome size variation in *Zea mays* ssp. *mays* adapted to different altitudes. Theor. Appl. Genet. 79, 470–474.

Rees, H., Durrant, A., 1986. Recombination and genome size. Theor. Appl. Genet. 73, 72–76.

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, 192–203.

Roth, G., Nishikawa, K.C., Wake, D.B., 1997. Genome size, secondary simplification, and the evolution of the brain in salamanders. Brain Behav. Evol. 50, 50–59.

SanMiguel, P., Bennetzen, J.L., 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. 82 (Suppl. A), 37–44.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L., 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20, 43–45.

Schlötterer, C., Harr, B., 2000. *Drosophila virilis* has long and highly polymorphic microsatellites. Mol. Biol. Evol. 17, 1641–1646.

Selosse, M.-A., Albert, B., Godelle, B., 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. 16, 135–141.

Semple, C., Wolfe, K.H., 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. 48, 555–564.

Shirasu, K., Schulman, A.H., Lahaye, T., Schulze-Lefert, P., 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. 10, 908–915.

Shuter, B.J., Thomas, J.E., Taylor, W.D., Zimmerman, A.M., 1983. Phenotypic correlates of genomic DNA content in unicellular eukaryotes and other cells. Am. Nat. 122, 26–44.

Simillion, C., Vanepoele, K., Van Montagu, M.C.E., Zabeau, M., Van de Peer, Y., 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U. S. A. 99, 13627–13632.

Stepkowski, T., Legocki, A.B., 2001. Reduction of bacterial genome size and expansion resulting from obligate intracellular lifestyle and adaptation to soil habitat. Acta Biochim. Pol. 48, 367–381.

Thomas, C.A., 1971. The genetic organization of chromosomes. Annu. Rev. Genet. 5, 237–256.

Thomson, K.S., Muraszko, K., 1978. Estimation of cell size and DNA content in fossil fishes and amphibians. J. Exp. Zool. 205, 315–320.

Turpeinen, T., Kulmala, J., Nevo, E., 1999. Genome size variation in *Hordeum spontaneum* populations. Genome 42, 1094–1099.

Vicient, C.M., Schulman, A.H., 2002. *Copia*-like retrotransposons in the rice genome: few and assorted. Genome Lett. 1, 35–47.

Vicient, C.M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharev, A., Nevo, E., Schulman, A.H., 1999a. Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell 11, 1769–1784.

Vicient, C.M., Kalendar, R., Anamthawat-Jónsson, K., Suoniemi, A., Schulman, A.H., 1999b. Structure, functionality, and evolution of the *BARE*-1 retrotransposon of barley. Genetica 107, 53–63.

Vieira, C., Nardon, C., Arpin, C., Lepetit, D., Biémont, C., 2002. Evolution of genome size in *Drosophila*. Is the invader's genome being invaded by transposable elements? Mol. Biol. Evol. 19, 1154–1161.

Vinogradov, A.E., 1995. Nucleotypic effect in homeotherms: body mass-corrected basal metabolic rate of mammals is related to genome size. Evolution 49, 1249–1259.

Vitte, C., Panaud, O., 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol. Biol. Evol. 20, 528–540.

Vinogradov, A.E., 2002. Growth and decline of introns. Trends Genet. 18, 232–236.

Walbot, V., 1999. UV-B damage amplified by transposons in maize. Nature 397, 398–399.

Waltari, E., Edwards, S.V., 2002. Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. Am. Nat. 160, 539–552.

Watanabe, K., Yahara, T., Denda, T., Kosuge, K., 1999. Chromosomal evolution in the genus *Brachyscome* (Asteraceae, Astereae): statistical tests regarding correlation between changes in karyotype and habit using phylogenetic information. J. Plant Res. 112, 145–161.

Wendel, J.F., Cronn, R.C., Johnston, J.S., Price, H.J., 2002a. Feast and famine in plant genomes. Genetica 115, 37–47.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L., Senchina, D.S., 2002b. Intron size and genome size in plants. Mol. Biol. Evol. 19, 2346–2352.

Wickham, S.A., Lynn, D.H., 1990. Relations between growth rate, cell size, and DNA content in colpodean ciliates (Ciliophora: Colpodea). Eur. J. Protistol. 25, 345–352.

Wu, C.-I., Li, W.-H., 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. U. S. A. 82, 1741–1745.