OPINION

Molecular melodies in high and low *C*

Daniel L. Hartl

For 50 years now, one of the enigmas of molecular evolution has been the *C*-value paradox, which refers to the often massive, counterintuitive and seemingly arbitrary differences in genome size observed among eukaryotic organisms. For example, the genome of the fruitfly *Drosophila melanogaster* is 180 megabases (Mb), whereas that of the European brown grasshopper *Podisma pedestris* is 18,000 Mb. The difference in genome size of a factor of 100 is difficult to explain in view of the apparently similar levels of evolutionary, developmental and behavioural complexity of these organisms.

The *C*-value paradox emerged from among the first applications of spectrophotometric analysis of nuclear DNA content[1]. The haploid DNA content of eukaryotic organisms ranges over a factor of 80,000. Some of the largest genomes are found among the lowliest of eukaryotes, such as the amoebae, and some of the smallest genomes are found among organisms with complex developmental and behavioural repertoires, such as *Drosophila melanogaster*. These discoveries were made before the elucidation of the molecular structure of DNA or its genetic coding function, so it is understandable that massive differences in DNA content were difficult to interpret. In the subsequent two decades molecular biologists laid out the molecular mechanistic framework of life — replication, transcription, translation and mutation. But at the culmination of this period, the *C*-value paradox was as great a mystery as ever. Maybe the paradox lay within ourselves. What if our concepts of organismic complexity were backwards? Perhaps the lower forms actually do have more genes — maybe, in fact, "they require more genes to conduct their dreary affairs"[2].

DNA renaturation kinetics carried out on many eukaryotes showed that genomic DNA contains many moderately or highly repetitive sequences, the relative amounts of which can differ markedly from one species to the next[3,4]. Many of the differences in genome size can be attributed to differences in the abundance of these repetitive sequences, rather than to large differences in the non-repetitive fraction of unique DNA, which includes the coding sequences[5]. Large-scale genomic sequencing gives a quantitative picture. On the long arm of human chromosome 22 (REF. 6), only 39 per cent of the DNA sequence resides in annotated genes, including their introns, and only three per cent resides in the exons of the annotated genes; in contrast, about 42 per cent of the chromosome consists of tandem and interspersed repeats of various kinds, including 16.8 per cent *Alu* repeats, 9.7 per cent *LINE 1* repeats, and 3.8 per cent *LINE 2* repeats. On chromosome 21 the situation is similar, but with only 26.2 per cent of the DNA in annotated genes[7]. To a large extent the *C*-value paradox is due to the proliferation or diminution of repetitive elements.

The players

Some of the main mechanisms for change in genome size are shown in FIG. 1. We include chromosomal mechanisms, such as polyploidy and accessory chromosomes, even though these mechanisms are prominent only in certain lineages, particularly in plants.

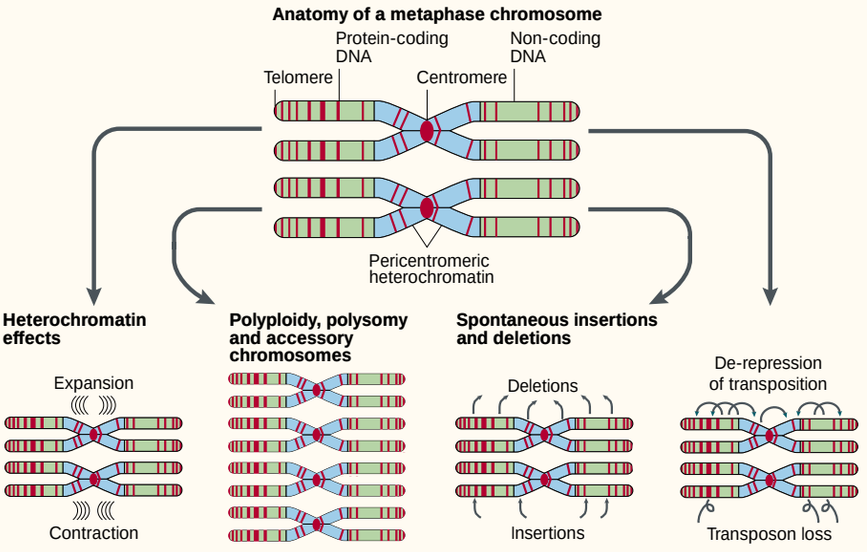


Figure 1 | **Principal mechanisms for changes in genome size.** In a large genome, such as the human genome, the protein-coding DNA is sparse and interspersed with non-coding DNA; at the scale shown here, coding DNA would be invisible. Except in some plant lineages, polyploidy is not a principal cause of variation in genome size. Insertions and deletions differ in size as well as in rate among species of organisms.

In some lineages in which polyploidy does take place, most of the differences in genome size in different species are nevertheless due to other causes. For example, the fact that wheat (genome size 16,000 Mb) is hexaploid accounts for only about 8 per cent of its genome size relative to that of rice (genome size 430 Mb), because the wheat genomes contain large amounts of repetitive DNA that are not present in the rice genome.

FIGURE 1 focuses on the mutational mechanisms that can change genome size, but natural selection may act on the genetic variation created by mutation. With regard to selection for genome size, there is an extensive literature on potential adaptive functions of non-coding DNA, much of it related to correlations between genome size and cellular traits (notably nuclear volume) or organismic traits (notably developmental time)[8]. Amoebas, with among the largest genomes, also have among the largest cells; in describing an entamoebal infection in 1890, William Osler[9] observed:

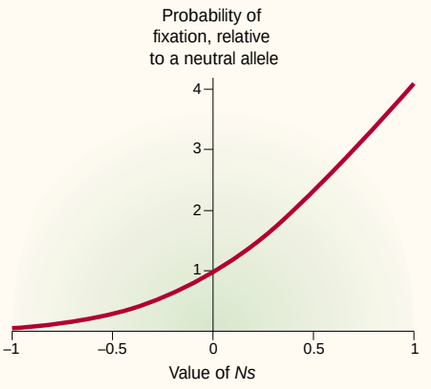
> "They are most extraordinary and striking creatures and take one's breath away at first to see these big amoebae — 10–20 times the size of a leucocyte — crawling about in the pus."

Limitations of space preclude an extensive discussion here, but the varieties of adaptive hypotheses for the maintenance of non-coding DNA include the 'skeletal DNA' hypothesis[10], according to which non-coding DNA functions as part of the basic framework for the assembly of the nucleus and serves to regulate nuclear volume in relation to cell volume; and the 'buffering DNA' hypothesis[11], which posits that non-coding DNA buffers condensed chromatin from intracellular solutes, and uncondensed chromatin from nonspecific DNA binding by proteins and other ligands. Conversely, views of non-coding DNA as merely accumulated 'junk DNA'[12] or self-perpetuating 'selfish DNA'[13,14] stand against these adaptionist models of genome evolution. Recent evidence showing that non-coding DNA is subject to elimination comes from studies of cryptomonads and chlorarachneans[15,16]. In these organisms, the descendants of ancient symbioses, the nucleus of a former algal partner persists as a simplified 'nucleomorph', surrounded by a periplastid membrane; in different lineages, the nucleomorph has undergone a 200–1,000-fold reduction in genome size with the elimination of virtually all of the non-coding DNA.

Perhaps lower forms have more genes — maybe, in fact, "they require more genes to conduct their dreary affairs".

Our present focus is on changes in genome size that are individually small relative to the total genome, so the points that we wish to make stand apart from discussions of the potential functional role or roles of non-coding DNA and possible mechanisms of selection for overall genome size. The magnitude of the changes is relevant because it is entirely possible for overall genome size to be a target of selection, but for small individual changes to be nearly neutral. An analogy may help to make this clear. It is known that the genome of warm-blooded vertebrates consists of megabase-long stretches of relatively G+C-rich or G+C-poor 'isochores', which are almost certainly maintained by selection[17]. But selection for overall G+C content of an isochore does not preclude a G–C base pair being replaced with a C–G base pair, or even an occasional A–T base pair becoming fixed by chance, as it is the overall G+C content that is critical rather than each individual base pair. So also with genome size: there may well be selection for the aggregate genome size, but insertions or deletions that are sufficiently small virtually escape selective forces, except in the aggregate. The figure in BOX 1 shows the quantitative relationship between the probability of fixation of a new mutant allele, relative to a neutral allele, and the product of the selective effect of the allele (s) and the effective population size (N). The theoretical basis of this curve is explained in BOX 1. It indicates that, for changes in genome size that are small relative to the total genome size, the evolutionary process governing genome size involves a balance between mutation, weak selection and random GENETIC DRIFT[18].

Box 1 | The fate of small differences in genome size



The efficacy of selection acting on a mutant allele is inevitably limited by the selective effect, s, of the mutant allele itself, as well as by the effective population size, N. The ultimate fate of a new mutant allele whose selective effect, positive or negative, satisfies $|Ns| << 1$ is subject to random genetic drift. Therefore, a deleterious allele can become fixed in spite of selection against it, or a beneficial allele can become lost in spite of selection favouring it. The curve shown here is based on a formula originally derived by Kimura[51], which asserts that the probability of fixation of a new mutant allele is roughly given by $2s/(1-e^{-4Ns})$, where s is the selective advantage (if $s > 0$) or disadvantage (if $s < 0$) of the allele, and N is the effective population size. This formula assumes that the original mutation occurs in a single individual, so that its initial frequency in the population is $1/2N$. For a new mutant allele that is neutral in its selective effects ($s = 0$), the probability of ultimate fixation is $1/2N$. So the probability of fixation of any new mutant allele, relative to that of a neutral mutation, is $4Ns/(1-e^{-4Ns})$, which has the convenient feature of depending only on Ns. This is the curve illustrated. Note that a deleterious allele with $Ns = -0.5$ still has a chance of becoming fixed that is 31.3 per cent of that of a neutral allele; in contrast, a beneficial allele with $Ns = +0.4$ has only twice the chance of becoming fixed as a neutral allele.

Scattered grace notes

Some of the mechanisms in FIG. 1 have effects that are restricted to particular nuclear compartments or classes of DNA. These include cytogenetic mechanisms such as the accumulation of accessory chromosomes, which are usually composed largely of heterochromatin. They also include as yet poorly understood mechanisms that result in the expansion or contraction of the amount of satellite DNA, composed of tandemly repeated simple sequences located primarily in the pericentromeric heterochromatin, or the expansion or diminution of the amount of pericentromeric heterochromatin itself[19,20]. Whereas these mechanisms act locally in the genome, other mechanisms act globally, although not necessarily to the same extent in all nuclear compartments or classes of DNA. Included here are insertions of transposable elements, as well as insertions and deletions of various sizes, resulting from processes unrelated to transposition.

In the evolution of genome size, both local and global mechanisms are at work. Compare, for example, *D. melanogaster* with *D. virilis*, which are estimated to have diverged from a common ancestor 40–60 million years

ago[21]. The genome size of *D. melanogaster* is 180 Mb, that of *D. virilis* about 300 Mb (REF. 5), but this difference is unequally apportioned between euchromatin and heterochromatin. In *D. melanogaster*, the euchromatin consists of 120 Mb and the pericentromeric heterochromatin 60 Mb (REF. 22), whereas in *D. virilis* the comparable figures are both estimated at 150 Mb (REF. 5). So the euchromatic genome of *D. virilis* is 25 per cent larger than that of *D. melanogaster*, whereas the heterochromatic genome is 150 per cent larger.

On the other hand, differences in genome size affecting the euchromatin seem to result from a large number of relatively small changes scattered throughout the euchromatin. Supporting evidence for this assertion consists of a positive correlation between the average size of orthologous introns and genome size between species of mammals[23] as well as between species of *Drosophila*[24]. But extensive comparisons of intron size among ten eukaryotic model organisms indicate that the correlation between intron size and genome size, although significant, is weak[25], and across a wide evolutionary range the difference in intron size is smaller than the difference in genome size by four orders of magnitude[26].

Tempo: presto, andante and adagio

Some of the mechanisms in FIG. 1 act on vastly different timescales. Polyploidy is at the faster end of the spectrum, bringing about twofold to fourfold increases in genome size over a single generation. High activity of transposable elements can also increase genome size by substantial amounts over a relatively short period[27]. Changes in genome size through small spontaneous deletions and insertions are at the slower end of the spectrum; although *Drosophila* has a relatively rapid rate of DNA loss through the fixation of spontaneous deletions, the absolute rate of DNA loss in *D. melanogaster* through this process is slow — about one base pair per generation[28]. Over the long haul, slow but steady forces may be as important in establishing the equilibrium genome size as forces that have large effects over short intervals, but it is generally unknown whether the genome size of any particular species is, in fact, at an equilibrium. The realized genome size at any time may result from a dynamic and constantly shifting balance between slow deletion and rapid transposon proliferation[29].

Transpositions of scale

The proliferation of transposable elements provides a mechanism for rapid increase in genome size. Comparison of a 240 kb region around the *alcohol dehydrogenase 1* (*adh1*) gene of maize and sorghum has indicated that this region of the maize genome has grown significantly in size in the 16 million years since these species last shared a common ancestor. In maize, the *adh1* region is composed of 33–62 per cent high-copy-number retrotransposons from 11 families (retrotransposons transpose using an RNA intermediate) and 16 per cent middle- and low-copy-number retrotransposons, whereas none of these insertions is found in the corresponding region of the sorghum genome[27].

More precise estimates of the timescale of maize genome expansion have been obtained by comparing the long terminal repeats (LTRs) found at the ends of LTR-containing retrotransposons, which become identical in nucleotide sequence during the process of transposition[30]. After transposition, the LTRs are free to diverge in sequence, and the time of each insertion can be estimated from the extent of the divergence. Analysis of 17 retrotransposons in the *adh1* region and two from other regions implies that, over the past three million years, the high rate of fixation of retrotransposons has increased the size of the maize genome from 1,200 Mb to 2,400 Mb (REF. 30). Although the data do not distinguish between an increased rate of transposition and an increased probability of fixation resulting from natural selection for a larger genome, the former interpretation seems more probable.

Unregulated transposition results in an exponential increase in the abundance of a family of transposons. However, for most transposable elements, the accumulation of copies in a genome is limited by regulatory factors encoded in the element itself, in the host, or through their interactions[31–35]. Quantitative aspects of transposon regulation relative to effects on genome size are summarized in BOX 2. Forms of regulation that are continuous functions of copy number may also be overridden by co-suppression and related phenomena[36,37], which can shut down transposition almost completely[38–40]. For example, in one group of South American rodents, including rice rats of the genus *Oryzomys*, the *LINE 1* elements that are prevalent in other mammalian genomes seem to have been quiescent for so long that they may have been eliminated from the genome[41].

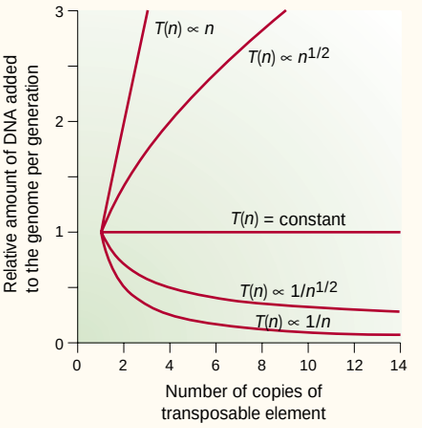
Insertions and deletions

The spectrum of spontaneous insertions and deletions (indels) is one of the relevant parameters in the long-term evolution of genome size. By the 'spectrum' of deletions or insertions we mean the product of frequency and average size. The effects, slow but persistent, are seemingly inescapable. If the spectrum of spontaneous deletions were greater than that of spontaneous insertions, this would create an ineluctable global mutation pressure toward a smaller genome size. The opposite would be true if the spectrum of deletions were smaller than that of insertions. Over evolutionary time, differences in the indel spectra among lineages may result in differences in genome size.

Mammalian PSEUDOGENES were the first sequences whose indel spectrum was analysed in the context of genome size[42].

Box 2 | Regulation of transposition

For most transposable elements, regulatory factors decease the rate of transposition per copy as the copy number increases. The mathematical form of the relationship is difficult to specify, and seems to differ from one transposon to the next[52,53]. Some simple forms of regulation are shown by the accompanying curves[53]. In these curves, n is the copy number per genome, and $T(n)$ is the expected increase in copy number in one generation when n copies are present, so the rate of transposition *per copy* is $T(n)/n$. For a family of transposons regulated according to the function $T(n)$, the increase in genome size per generation resulting from transposition is given by $n' - n = T(n)$, disregarding any possible selection for larger or smaller genome size. With no regulation, $T(n)$ increases linearly in n, and the increase in abundance of the transposon is exponential. Any form of regulation slows the rate of increase. For example, if the overall rate of transposition $T(n)$ is a constant and independent of copy number, then the increase in genome size resulting from transposition is also a constant that depends on the length of the transposon. Stronger types of regulation diminish the rate of increase in genome size further.



Spontaneous deletions outnumbered insertions and had a longer mean length. The rate of DNA loss was estimated to be faster in the mouse lineage than in the human lineage, consistent with the smaller size of the mouse genome, but the overall rate of DNA loss seemed to be almost negligible.

Analysis of non-functional sequences, such as pseudogenes in the mammalian genome, is essential to avoid bias in the estimation of mutation rates, including those of indels, from sequence data. This is because the probability of fixation of new mutations is proportional to the mutation rate only for non-functional sequences that are not subject to selective constraints. All functional sequences are constrained, to some extent, by selection, and so the proportionality breaks down. Unfortunately, pseudogenes seem to be widespread only in organisms with relatively large genomes, and so pseudogenes have limited applicability to studying spontaneous mutations and the indel spectra in diverse organisms.

But eukaryotes contain classes of non-functional DNA other than pseudogenes. One candidate consists of non-LTR retrotransposable elements, which are ubiquitous in eukaryotes with the exception of yeast. Their molecular mechanism of transposition generates a predominance of non-functional 'dead-on-arrival' (DOA) copies. Because the processivity of the reverse transcriptase of non-LTR retrotransposons is substantially shorter than the total length of the elements, most new copies lack the promoter end and so cannot mobilize themselves, and only very rarely can they be mobilized *in trans*. For such elements, molecular phylogenetic analysis can be used to separate mutations that occur in active, master lineages, which are subject to purifying selection, from those that occur in the inactive, pseudogene-like DOA copies[28,43,44]. The set of mutations in DOA copies, because they are not individually targets of selection, can then be used to estimate the patterns of spontaneous point mutation and indel formation.

Glossary

GENETIC DRIFT
Random changes in allele frequency that result because the genes appearing in offspring are not a perfectly representative sampling of the parental genes (e.g. in small populations).

PSEUDOGENE
A DNA sequence originally derived from a functional protein-coding gene that has lost its function owing to the presence of one or more inactivating mutations.

ORTHOLOGOUS GENES
Homologous genes in different species whose lineages derive from a common ancestral gene without gene duplication or horizontal transmission.

Figure 2 | **The mountain grasshopper *Podisma pedestris*.** This species has a genome size of 18,000 Mb, tenfold larger than the *Drosophila* genome and sixfold larger than the human genome. (Courtesy of G. Hewitt.)

Because non-LTR elements are widespread in eukaryotes and can be cloned relatively easily[45,46], this method can be used to study mutational patterns in a diversity of organisms. In principle, it can distinguish the evolutionary dynamics of different genomic compartments, such as euchromatin and heterochromatin, by choosing non-LTR elements that target these compartments.

The initial studies of DOA elements in *Drosophila* revealed a striking difference in the indel spectra between *Drosophila* and mammals. We found a 60-fold faster rate of DNA loss in *Drosophila*, consistent with its much more compact genome[28,43]. To test whether the apparently high rate of DNA loss in *Drosophila* could be due to deletions introduced at the time of transposition, we examined recently transposed copies of the transposon, which showed no enhanced rates of mutations of any kind[18]. Consistent with the compact genome of *Caenorhabditis elegans*, analysis of the large *srh* family of chemoreceptor genes has revealed a rapid rate of DNA loss comparable to that in *Drosophila*[47]. On the other hand, analysis of a non-LTR retrotransposon isolated from Hawaiian crickets of the genus *Laupala*, which have a genome that is 11-fold larger than that of *Drosophila*, showed a 40-fold slower rate of DNA loss than that observed in *Drosophila*[46]. Similarly, nuclear pseudogenes derived from mitochondrial DNA in the European brown grasshopper *Podisma pedestris* have a rate of DNA loss even slower than that in *Laupala*, in agreement with its approximately tenfold larger genome size[48] (FIG. 2).

These results indicate that differences in indel spectra may underlie (or, in any event, are consistent with) at least some important differences in genome size. How widely this correlation will hold remains to be determined. However, the findings do not preclude selection acting on genome size as a whole. In principle, the differences in the indel spectra observed among organisms could be due either to differences in patterns of spontaneous mutation or to differences in natural selection for overall genome size affecting the fate of individual indels[49]. We have argued on general theoretical grounds that the fixation of individual indels, whose size is small in comparison with that of the total genome, would be expected to be governed primarily by random genetic drift[18,50]. Experimental studies of the fate of indels in different lineages, or of the allele-frequency spectra of polymorphic indels in unconstrained sequences, would be welcome to address this issue in greater depth.

Finale

Selection for overall genome size is expected to be relatively ineffective in influencing the fate of individual changes in genome size that are sufficiently small relative to the total. So molecular studies of such changes can be used to infer the underlying mutational processes affecting genome size.

The euchromatic genome often seems to change in size through relatively small insertions or deletions that are more or less uniformly distributed. This pattern is seen in the correlations of intron sizes in ORTHOLOGOUS GENES among related species that differ in genome size[23,24]. Transposable elements make up a significant part of the euchromatin in some species, and de-repressed transposition can change genome size rapidly, leading, in the case of maize, to an estimated doubling in genome size in the past three million years[30]. Genomes can also increase or decrease in size by the random fixation of spontaneous insertions or deletions in non-essential DNA, but this is a much slower process. Nevertheless, there is a significant inverse correlation between rate of DNA loss and overall genome size in mammals[42] and in insects[46]. Although such studies have revealed a great deal about the mutational components of change in genome size, the inference of near-neutrality of small changes does not preclude natural selection acting on the overall genome size through such mechanisms as maintaining nuclear volume. Analogously, selection for overall G+C content in isochores need not be effective in maintaining the identity of each individual base pair.

Daniel L. Hartl is at the Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. e-mail: dhartl@oeb.harvard.edu

Links

FURTHER INFORMATION Hartl lab page
ENCYCLOPEDIA OF LIFE SCIENCES Genome organization/ humans

1. Mirsky, A. E. & Ris, H. The DNA content of animal cells and its evolutionary significance. *J. Gen. Physiol.* **34**, 451–462 (1951).
2. Thomas, C. A. The genetic organization of chromosomes. *Annu. Rev. Genet.* **5**, 237–256 (1971).
3. Bonner, J. *et al.* Functional organization of the mammalian genome. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 303–310 (1973).
4. Davidson, E. H., Hough, B. R., Amenson, C. S. & Britten, R. J. General interspersion of repetitive with nonrepetitive sequence elements in the DNA of *Xenopus*. *J. Mol. Biol.* **77**, 1–23 (1973).
5. John, B. & Miklos, G. L. G. *The Eukaryotic Genome in Development and Evolution* 1–416 (Allen and Unwin, London, 1988).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
8. Cavalier–Smith, T. *The Evolution of Genome Size* (John Wiley, New York, 1985).
9. Bliss, M. *William Osler: A Life in Medicine* (Oxford, New York, 1999).
10. Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the C-value paradox. *J. Cell Sci.* **34**, 247–278 (1978).
11. Vinogradov, A. E. Buffering: A possible passive-homeostasis role for redundant DNA. *J. Theor. Biol.* **193**, 197–199 (1998).
12. Ohno, S. in *Evolution of Genetic Systems, Brookhaven Symp. Biol.* (ed. Smith, H. H.) 366–370 (1972).
13. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
14. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
15. Cavalier-Smith, T. & Beaton, M. J. The skeletal function of non-genic nuclear DNA: New evidence from ancient cell chimaeras. *Genetica* **106**, 3–13 (1999).
16. Beaton, M. J. & Cavalier-Smith, T. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. *Proc. R. Soc. Lond. B* **266**, 2053–2059 (1999).
17. Bernardi, G. Isochores and the evolutionary genetics of vertebrates. *Gene* **241**, 3–17 (2000).
18. Lozovskaya, E. R., Nurminsky, D. I., Petrov, D. A. & Hartl, D. L. Genome size as a mutation-selection-drift process. *Genes Genet. Syst.* **74**, 201–207 (1999).
19. Laurent, A. M., Puechberty, J. & Roizes, G. Hypothesis: for the worst and for the best, L1Hs retrotransposons actively participate in the evolution of the human centromeric alphoid sequences. *Chromosome Res.* **7**, 305–317 (1999).
20. Csink, A. K. & Henikoff, S. Something from nothing: The evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204 (1998).
21. Russo, C. A. M., Takezaki, N. & Nei, M. Molecular phylogeny and divergence times of Drosophilid species. *Mol. Biol. Evol.* **12**, 391–404 (1995).
22. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287**, 2185–2195 (2000).
23. Ogata, H., Fujibuchi, W. & Kanehisa, M. The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* **390**, 99–103 (1996).
24. Moriyama, E. N., Petrov, D. A. & Hartl, D. L. Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15**, 770–773 (1997).
25. Deutsch, M. & Long, M. Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–3228 (1999).
26. Vinogradov, A. E. Intron–genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49**, 376–384 (1999).
27. SanMiguel, P. & Bennetzen, J. L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* **82**, 37–44 (1998).
28. Petrov, D. A. & Hartl, D. L. High rate of DNA loss in the *D. melanogaster* and *D. virilis* species groups. *Mol. Biol. Evol.* **15**, 293–302 (1998).
29. Bennetzen, J. L. & Kellogg, E. A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**, 1509–1514 (1997).
30. San Miguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nature Genet.* **20**, 43–45 (1998).
31. Andrews, J. D. & Gloor, G. B. A role for the *KP* leucine zipper in regulating *P* element transposition in *Drosophila melanogaster*. *Genetics* **141**, 587–594 (1995).
32. Lohe, A. R. & Hartl, D. L. Autoregulation of *mariner* transposase activity by overproduction and dominant-negative complementation. *Mol. Biol. Evol.* **13**, 549–555 (1996).
33. Stellwagen, A. E. & Craig, N. L. Mobile DNA elements: controlling transposition with ATP-dependent molecular switches. *Trends Biochem. Sci.* **23**, 486–490 (1998).
34. Braam, L. A. M., Goryshin, I. Y. & Reznikoff, W. S. A mechanism for Tn*5* inhibition: Carboxyl-terminal dimerization. *J. Biol. Chem.* **274**, 86–92 (1999).
35. Sakai, J. S., Kleckner, N., Yang, X. & Guhathakurta, A. Tn*10* transpososome assembly involves a folded intermediate that must be unfolded for target capture and strand transfer. *EMBO J.* **19**, 776–785 (2000).
36. Bennetzen, J. L. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269 (2000).
37. Ketting, R. F. & Plasterk, R. H. A. A genetic link between co-suppression and RNA interference in *C. elegans*. *Nature* **404**, 296–298 (2000).
38. Chaboissiert, M. C., Bucheton, A. & Finnegan, D. J. Copy number control of a transposable element, the I factor, a LINE-like element in *Drosophila*. *Proc. Natl Acad. Sci. USA* **95**, 11781–11785 (1998).
39. Birchler, J. A., Pal-Bhadra, M. & Bhadra, U. Less from more: cosuppression of transposable elements. *Nature Genet.* **21**, 148–149 (1999).
40. Jensen, S., Gassama, M. P. & Heidmann, T. Cosuppression of *I* transposon activity in *Drosophila* by *I*-containing sense and antisense transgenes. *Genetics* **153**, 1767–1774 (1999).
41. Casavant, N. C. *et al.* The end of the LINE?: Lack of recent L1 activity in a group of South American rodents. *Genetics* **154**, 1809–1817 (2000).
42. Graur, D., Shuali, Y. & Li, W.-H. Deletions in processes pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* **28**, 279–285 (1989).
43. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1996).
44. Petrov, D. A. & Hartl, D. L. Trash DNA is what gets thrown away: High rate of DNA loss in *Drosophila*. *Gene* **205**, 279–289 (1997).
45. Wright, D. A. *et al.* Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*. *Genetics* **142**, 569–578 (1996).
46. Petrov, D. A., Sangster, T., Johnston, J. S., Hartl, D. L. & Shaw, K. L. Evidence for DNA loss as a determinant of genome size. *Science* **287**, 1060–1062 (2000).
47. Robertson, H. M. 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 (2000).
48. Bensasson, D., Petrov, D. A., Zhang, D. -X., Hartl, D. L. & Hewitt, G. M. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. Evol.* (in the press).
49. Charlesworth, B. The changing sizes of genes. *Nature* **384**, 315–316 (1996).
50. Petrov, D. A. & Hartl, D. L. Pseudogene evolution and natural selection for a compact genome. *Heredity* **91**, 221–227 (2000).
51. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719 (1962).
52. Simmons, M. J. & Bucholz, L. M. Transposase titration in *Drosophila melanogaster*: A model of cytotype in the P-M system of hybrid dysgenesis. *Proc. Natl Acad. Sci. USA* **82**, 8119–8123 (1985).
53. Sawyer, S. A. *et al.* Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics* **115**, 51–63 (1987).

Acknowledgements

I am very grateful to my colleagues D. Petrov and E. Lozovskaya for their contributions to the work described here, and to D. Petrov for having allowed me to read an early draft of a manuscript of his own. This work was supported by a grant from the National Institutes of Health.

TIMELINE

Environmental health and genomics: visions and implications

Kenneth Olden and Samuel Wilson

The relationship between genes and the environment can be compared to a loaded gun and its trigger. A loaded gun by itself causes no harm; it is only when the trigger is pulled that the potential for harm is released. Genetic susceptibility creates an analogous situation, where the loaded gun is one or a combination of susceptibility genes (alleles) and the trigger is an environmental exposure. The key objective of the Environmental Genome Project is to identify alleles that confer susceptibility to the adverse effects of environmental agents. Here we discuss the goals of the Environmental Genome Project, its implications and, in particular, its potential effect on our ability to assess human disease risk in the future.

Scientists in biomedicine, environmental health and public health are working to understand and prevent human disease. The identification and functional characterization of susceptibility genes is critical to achieving this goal and for predicting risk from environmental exposure and response to pharmaceuticals. Many chronic diseases in humans arise from a complex array of factors, which could include several genes, environmental conditions or exposures, the age, nutritional status or stage of development of a person, and other predisposing factors. Therefore, most chronic diseases will not be fully understood until both the genetic and environmental contributions to their aetiology are elucidated. Unfortunately, the relationship between genes and the environment is neither well