Synergy between sequence and size in Large-scale genomics

Article *in* Nature Reviews Genetics · October 2005

DOI: 10.1038/nrg1674 · Source: PubMed

1 author:

T. Ryan Gregory
University of Guelph
106 PUBLICATIONS **6,379** CITATIONS

SEE PROFILE

SYNERGY BETWEEN SEQUENCE AND SIZE IN LARGE-SCALE GENOMICS

T. Ryan Gregory

Abstract | Until recently the study of individual DNA sequences and of total DNA content (the C-value) sat at opposite ends of the spectrum in genome biology. For gene sequencers, the vast stretches of non-coding DNA found in eukaryotic genomes were largely considered to be an annoyance, whereas genome-size researchers attributed little relevance to specific nucleotide sequences. However, the dawn of comprehensive genome sequencing has allowed a new synergy between these fields, with sequence data providing novel insights into genome-size evolution, and with genome-size data being of both practical and theoretical significance for large-scale sequence analysis. In combination, these formerly disconnected disciplines are poised to deliver a greatly improved understanding of genome structure and evolution.

Because it has unfolded almost entirely within the past decade, the history of complete genome sequencing is generally well known among contemporary researchers. By contrast, the other form of large-scale genomics — the study of total genome size (in terms of mass or base pairs) — persists as one of the longest-running puzzles in genetics, even pre-dating the demonstration of DNA as the hereditary material and the elucidation of its molecular structure. In fact, the constancy of haploid nuclear DNA amounts (C-values) within individual organisms and species, first reported by Boivin *et al.* in 1948 (REF. 1), was taken as evidence that DNA, and not proteins, must be the substance of which genes are composed. But only a few years later, broader surveys of genome-size variation in animals[2] exposed a startling discrepancy between DNA content and organismal complexity (considered a proxy for gene number), an observation sufficiently perplexing to become known as the 'C-value paradox' two decades later[3] (BOX 1).

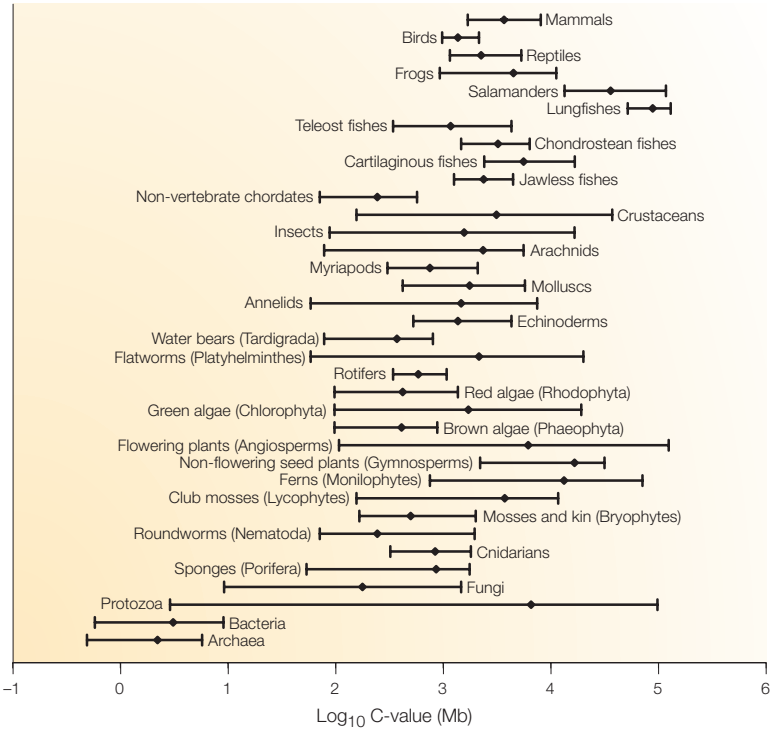
The discordance between genome size and organism complexity or gene number did not remain a paradox — that is, a pair of mutually exclusive truths — for very long. The discovery of non-coding DNA in the early 1970s explained the failure of DNA content to reflect the number of genes, and in so doing resolved the paradox. However, as with most significant advances in genetic knowledge, this finding raised more questions than it answered. The most prominent among these relate to the nature of non-coding DNA, which even in the first decade after its discovery was variously described as being 'junk' (that is, now-functionless gene copies, or 'pseudogenes'[4,5]), as serving a structural (nucleoskeletal) function[6], as consisting entirely of introns[7], and of representing strictly 'selfish' elements[8,9].

Although much has been learned over the past half-century of study into genome size, important questions remain to this day. For example, what types of sequence make up this non-coding majority, and in what proportions? How are these sequences gained and lost from genomes over population genetic and long-term evolutionary timescales? What effects, or perhaps even functions, if any, does this non-coding DNA have with respect to cellular and organismal phenotypes? Why are the chromosomes of some organisms (for example, birds) so gaunt, whereas those of others (for example, salamanders) are positively bloated? Unlike the former C-value paradox, the more complex 'C-value enigma' (as the sum of these puzzles is more appropriately called) has for decades defied all attempts at one-dimensional explanations[10]. It is now clear that the

Department of Integrative Biology, University of Guelph, Ontario N1G 2W1, Canada.
e-mail: rgregory@genomesize.com
doi:10.1038/nrg1674

Box 1 | **Extensive variation in genome size within and among the main groups of life**

Ever since the first general surveys of nuclear DNA content were carried out in the early 1950s it has been apparent that eukaryotic genome sizes vary enormously and that this is unrelated to intuitive ideas of morphological complexity[2]. This discrepancy between genome size and complexity remains clear more than half a century later, with genome sizes now available for nearly 9,000 species of animals and plants[10,11]. In prokaryotes, genome size and gene number are strongly correlated[86], but in eukaryotes the vast majority of nuclear DNA is non-coding (FIG. 1; BOX 3)**. Nevertheless, there is some overlap in genome size between the largest bacteria and the smallest parasitic protists. The figure illustrates the means and overall ranges of genome size that have been observed so far in the main groups of living organisms, and are loosely arranged according to common ideas of complexity to further emphasize the disparity between this parameter and genome size. Some commonly cited extreme values for amoebae (700,000 Mb) have been omitted, as there is considerable uncertainty about the accuracy of these measurements and the ploidy level of the species involved[10,87].**



C-value enigma will require the integration of insights derived from various disciplines including cytogenetics, cell biology, morphology, developmental biology, physiology, evolutionary theory, phylogenetics, ecology (BOX 2) and, as argued here, complete genome sequencing.

A detailed review of either genome sequencing or genome size is neither the intent nor within the scope of this discussion (for this, see REFS 10–12). Instead, the following sections outline some crucial new insights into the study of genome size that have been derived from complete sequences, and the importance of genome size in the generation and interpretation of genome sequences. The key message throughout this article is that considerable benefits are to be had by bridging the current divide between sequence and size.

Using sequences to understand sizes

Most previous work on genome-size evolution has involved carrying out interspecific comparisons of total DNA content, mostly to the exclusion of gene-level analyses. In particular, the primary focus has been on correlating variation in DNA content with a range of parameters, from the sizes of individual chromosomes to the geographical distribution of species[10,11,13–15] (BOX 2). Phenotypic associations such as these have had an important role in shaping discussions of genome-size evolution, but the obvious problem is that they deal only with the subset of the C-value enigma that relates to the implications of DNA-content variation. The equally important components of the puzzle that involve the sub-genomic processes and specific sequences that generate variation in genome size have received less attention. For the most part, this is because these issues can only be examined in detail through large-scale comparisons of DNA sequences, an approach that has become possible only relatively recently.

Fortunately, interest in the molecular bases of genome-size change has been increasing steadily over the past 10 years. This has included not only rudimentary analyses of the sequences and processes that add to genomic bulk, but also of previously overlooked mechanisms for genome shrinkage. The net result has been a recognition that genome sizes can change — in either direction — by various processes that operate at many physical and temporal scales, from individual replication events within genomes to filtering at the level of populations and higher-order lineages[10,15] (BOX 2). Some specific contributions of large-scale sequencing to this new understanding of genome-size change are highlighted in the following sections. A few warnings are also provided in an effort to prevent an overextension of these valuable, but still limited, genome-sequence data.

Box 2 | **Dealing with the enigma of genome size**

Several lines of inquiry have been used to shed light on the evolution of the large-scale features and key components of genomes, the most prominent of which are outlined here.

Cytogenetics and molecular biology
Complete genome sequencing provides the most comprehensive information about genomic components (BOX 3), but mostly remains restricted to species that have small genomes. Biochemical techniques to characterize repetitive fractions, cytogenetic analyses to examine chromatin condensation patterns, and other molecular methods for determining approximate copy numbers of different sequences have also been important in the study of genome size.

Cell biology, morphology, physiology, developmental biology and ecology
Correlations have been identified between DNA content and many features that range from the subcellular to the supraspecific: chromosome size; nucleus size; cell size; cell division rate; seed, pollen or egg size; body size; oxygen consumption or photosynthetic rate; developmental rate and/or the presence or intensity of metamorphosis; geographical distribution; and extinction risk[10,11,15]. Some of these patterns are universal, but others vary from group to group, which strongly indicates the importance of organismal biology in affecting genome-size change (and *vice versa*).

The study of mutational mechanisms
Increased attention has been paid in recent years to the mechanistic bases of changes in genome size, including a heightened emphasis on DNA loss. Models that relate to small insertion–deletion biases[49,69], deletion that is due to recombination among LTR retrotransposons[88,89], and variation in the efficacy of DNA repair[90,91] have all recently been explored. Duplications of genes and genomes, both recent and ancient, are also increasingly common subjects of study[55,92].

Population genetics
Population genetics models have long been used to explore the abstract conditions (for example, the reproductive mode, population size or environment of their hosts) under which transposable elements tend to flourish or falter[28,93–100], and therefore contribute to the broader understanding of how and why a given genome comes to accumulate the amount of DNA contained within it.

Phylogenetics and fossils
Historical patterns of genome-size change were long the subject of speculation, but an increased emphasis on phylogenetic analyses has provided clear insights into both directionality and timing. For example, analyses at small[101] and large[11,102] scales have indicated that both increases and decreases in genome size have occurred. Such studies also shed light on the timescales over which genome sizes have changed, information that had previously been available only in a small number of studies that examined fossil cell sizes[103–106].

Transposable elements and their hosts. The human genome sequence revealed, for the first time, not only which general categories of sequence, but also which specific elements contribute prominently to a relatively large genome. By far the leading component in this regard proved to be transposable elements — or, most commonly, inactive remnants thereof — which represent ~45% of the sequence of the human genome (BOX 3). Amazingly, the two most prevalent elements alone, the short interspersed nuclear element (SINE) *Alu* (present in >10^6 copies) and the long interspersed nuclear element (LINE) *LINE1* (>5×10^5 copies), account for 11% and 17% of the human sequence, respectively[16] — a fact that is all the more remarkable given that a mere ~1.5% consists of protein-coding regions.

Transposable elements have been studied from various perspectives that can contribute to an understanding of their evolution and impact on the genome. These include the use of phylogenetic and comparative sequencing studies to determine their historical relationships and patterns of activity; predicting through population genetics modelling the conditions under which transposable elements will tend to spread most effectively; and studies into their roles as mutagens, as contributors to chromosome structure and organization, and as sources of significant new genic and regulatory diversity[17]. It has also been useful to take an holistic view of the "ecology of the genome"[18–20], according to which transposable elements might compete with one another for insertion sites and other resources[21,22], or might rely on each other (as 'parasites of parasites') for their transposition, as with the dependency of SINE elements on LINE elements[17]. Similarly, transposable elements are increasingly found to interact in complex ways with their host genomes — ranging along a continuum from parasitism to mutualism[18,19] — in some cases having been incorporated into the regulatory machinery of the genome[23,24] or being co-opted in the evolution of key organism-level functions such as immunity[25] and stress response[26]. (Of course no single function can account for genome-size variation in general; there is no reason to believe, for example, that the average salamander requires 5 to 15 times as much gene regulation as a typical bird or mammal[10].)

Transposable elements: abundance and diversity. As more genome sequences are completed, it is becoming clear that genome size and total transposable-element content are strongly correlated[27,28] (FIG. 1). This is particularly evident in organisms such as maize, in which a surge of transposable-element activity has led to a doubling of genome size in only a few million years[29]. However, the situation has proved to be more complex than first imagined under early theories that most non-coding DNA is simply 'selfish', and indeed many recent observations have been rather counter-intuitive on this basis. For example, despite their abundance, even the most common LINE elements, DNA transposons and LTR (long terminal repeat) retrotransposons are effectively extinct or nearly so in the human genome, leaving only fossils behind[16]. In fact, notwithstanding a burst of *Alu* activity about 40 million years ago (Mya), there has been a steady decline in transposable-element activity in the lineage leading to humans in the time since the mammalian radiation[16]. That said, it seems that transposable elements have been responsible for generating many polymorphisms in human populations[30].

It is interesting that the mouse genome, although it is ~14% smaller than that of humans, contains many more active transposable elements, including *LINE1*, four different SINE elements, and three classes of endogenous retrovirus[31]. Indeed, as the following examples show, a pattern of low abundance, high activity and extensive diversity of transposable elements might apply to many smaller eukaryotic genomes.

At least 40 different transposable-element families are represented by young, recently active elements in the pufferfish *Takifugu rubripes* (formerly known as *Fugu rubripes*), despite its genome being among the smallest in vertebrates. But even the most common type, the LINE element *Maui*, is present in only 6,400 copies[32]. In the second pufferfish to be sequenced, *Tetraodon nigroviridis*, only 4,000 transposable-element copies are found in total — but this still represents 73 different types of element[33]. Genomes that have a higher proportion of DNA transposons, such as in *Drosophila melanogaster* and *Arabidopsis thaliana*, contain elements of more recent origin that are derived from more families than in mammals[16]; this is explained by the fact that DNA transposons tend to be more short-lived and to spread by horizontal transfer. The genome of *D. melanogaster*, for example, contains about 130 different transposable-element families (including 25 non-LTR and 28 LTR families), all of which are younger than 20 million years (Myr) (REF. 34).

There is therefore convincing evidence that many smaller genomes contain a surprisingly high diversity of transposable-element families. It also now seems that the diversity of lineages within individual transposable-element families might be higher in some smaller genomes. In mammals, the abundant *LINE1* elements tend to be represented by a single lineage,

Box 3 | **The main components of eukaryotic genomes**

Protein-coding genes

Although most prokaryotic chromosomes consist almost entirely of protein-coding genes[86], such elements make up a small fraction of most eukaryotic genomes (see figure). As a prime example, the human genome might contain as few as 20,000 genes, comprising less than 1.5% of the total genome sequence[16,82].

Introns

Shortly after their discovery, the non-coding intervening sequences within coding genes (introns) were suggested to account for the pronounced discrepancy between gene number and genome size[7]. It has also recently been suggested that most non-coding DNA in animals (but not plants) is intronic, which would imply that most of the genome is transcribed even though protein-coding regions represent a tiny minority[107,108]. At the very least, introns were found to account for more than a quarter of the draft human sequence[16]. Over a broad taxonomic scale, intron size and genome size are positively correlated[109], although within genera a correlation might (for example, *Drosophila*[110]) or might not (for example, *Gossypium*[111]) be observed.

Pseudogenes

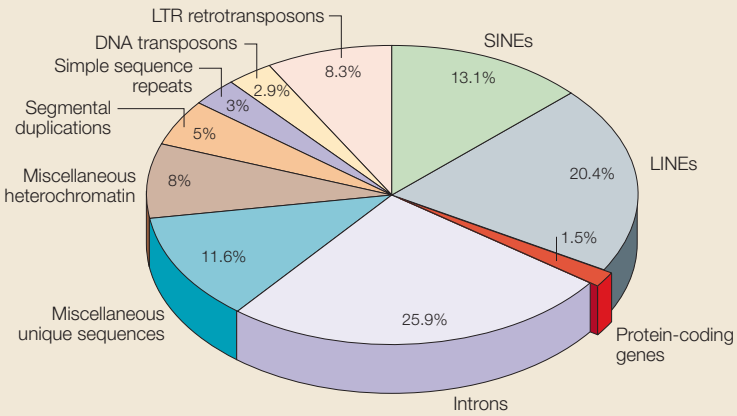
Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size[4]. However, it is now apparent that even in combination, 'classical pseudogenes' (direct DNA to DNA duplicates), 'processed pseudogenes' (copies that are reverse transcribed back into the genome from RNA and therefore lack introns) and 'Numts' (nuclear pseudogenes of mitochondrial origin) comprise a relatively small portion of mammalian genomes. The human genome is estimated to contain about 19,000 pseudogenes[46].

Transposable elements

In eukaryotes, transposable elements are divided into two general classes according to their mode of transposition. Class I elements transpose through an RNA intermediate. This class comprises long interspersed nuclear elements (LINEs), endogenous retroviruses, short interspersed nuclear elements (SINEs) and long terminal repeat (LTR) retrotransposons. Class II elements transpose directly from DNA to DNA, and include DNA transposons and miniature inverted repeat transposable elements (MITEs).

Transposable elements (and especially their extinct remnants) make up a large portion of the human genome, with some elements (for example, the SINE *Alu* element) present in more than a million copies. Transposable-element evolution involves complex interactions with the host genome and other subgenomic elements, ranging from parasitism to mutualism. For a review of transposable-element structure, origins, impacts and evolution see REF. 17.

The figure **provides a summary of the different components of the human genome. Less than 1.5% of the genome consists of the suspected 20,000–25,000 protein-coding sequences. By contrast, a large majority is made up of non-coding sequences such as introns (almost 26%) and (mostly defunct) transposable elements (nearly 45%). Data are taken from** REF. 16.



but the smaller genome of zebrafish contains more than 30 distinct lineages of this element[21]. In chickens, 85% of the constituent sequences cannot be identified, which probably reflects a high quantity of ancient transposable elements that have become degraded beyond recognition[35]. However, even here, several distinct lineages persist in the dominant chicken transposable element, the LINE element *CR1* (REF. 35).

Although it is probable that the overall abundance of each type of transposable element correlates positively with genome size[28], it is also evident that the proportion of element types varies tremendously among genomes. For example, LINE elements and SINE elements are dominant in mammals, but no active SINE elements have existed in avian genomes since the origin of birds[16,31,35,36]. LTR retrotransposons are the most common transposable elements in the grasses and in the silkworm moth *Bombyx mori*[37,38], whereas DNA transposons predominate in nematodes[39]. In the smallest genomes only a few — or in some cases no — transposable elements are found[40–45].

Lessons from the study of transposable elements. The evidence that has emerged from complete genome-sequencing projects indicates that small genomes are engaged in an active campaign to keep their diverse and active transposable-element populations in check, whereas in larger genomes one or a few types might spread relatively unhindered and then persist as identifiable fossils long after they lose their capabilities for self-replication. Conversely, it could be that organism-level selection for a small genome creates strong intragenomic selective pressure for the maintenance and/or diversification of active transposable elements. Certainly, deciphering the nature of the dynamics between the size and other features of host genomes and the genetic parasites of which many are largely composed represents a promising avenue for future investigation.

In short, although only a handful of eukaryotic genome sequences are currently available for comparison, it is already clear that the evolution of transposable elements — and by extension genome size — is a highly complex process that varies considerably in its specifics from one genome to the next, even among related organisms. Transposable elements might indeed propagate as 'selfish' elements with varying degrees of virulence and with diverse effects on genomes while they remain active, but once extinct (as the majority probably are in larger genomes) their abundance will depend on a complex interplay of mutational mechanisms, interactions with other elements, and evolutionary forces that are both internal and external to the genome (BOX 2).

Small-scale duplications. As with transposable elements, it seems that the total abundance of pseudogenes ('junk DNA', properly defined) might correlate positively with genome size, although overall these elements constitute a relatively small fraction of genome size. For example, only 51 pseudogenes are found in the chicken genome[35], 14,000 in the mouse[31], 18,755 in the rat[36] and 19,000 in the human[46], in accordance with their rankings with respect to genome size. The truly small-genomed species that have been sequenced so far all have low pseudogene numbers: only 33 are found in *Schizosaccharomyces pombe*[47], 166 in *Anopheles gambiae*, 176 in *D. melanogaster*[48] and 400 in *Oikopleura dioica*[41]. In keeping with this, there is evidence of a positive relationship between genome size and the estimated half-lives of new gene duplicates across species[28], further supporting the idea that a general propensity to delete non-coding DNA of all types correlates inversely with genome size[49,50].

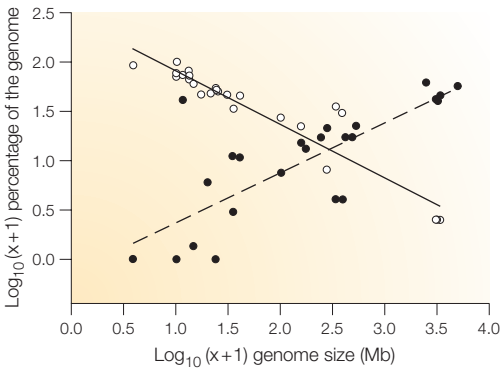


Figure 1 | **The relative contributions of two key components of eukaryotic genomes.** The relationships between haploid genome size and the percentage of the genome that consists of protein-coding genes (white circles) and transposable elements (black circles) are shown. The data are based on species that have been the subject of large-scale sequencing studies. Larger genomes contain proportionately fewer genes and more transposable elements than small genomes. A $\log_{10}(x+1)$ transformation was used because some tiny genomes contain no recognizable transposable elements.

Pseudogene evolution, similar to that of transposable elements, is being revealed by complete genome-sequence data as a complex process that can be influenced by several genomic factors. For example, there seem to be 'hot spots' of pseudogene formation near the centromeres of human chromosomes[46]. In mice and humans, the local abundance of processed pseudogenes seems to be linked to within-genome variation in GC content, as occurs with transposable elements[51]. Intriguingly, processed pseudogenes are especially rare in chickens, probably because the most prevalent chicken transposable element, *CR1* — unlike *LINE1* in humans — encodes a reverse transcriptase that is unlikely to copy polyadenylated mRNAs and therefore fails to generate them[35].

In terms of medium-scale duplication processes, it seems that more than 5% of the euchromatic human genome is composed of relatively recent segmental duplications (<40 Mya). It is interesting to note that the proportion of the rat genome (3%) that is made up of segmental duplications of at least 5 kb is intermediate between that of humans (5.3%) and mice (1–2%), which is in keeping with its intermediate genome size. This finding is consistent with the fact

that segmental duplications are smaller, less frequent and more likely to occur within a single chromosome in chicken versus mammals[35]. Again, such smaller-scale duplications might not strongly influence total DNA content, but their association with genome size hints at the existence of genome-wide patterns of DNA insertion and deletion irrespective of the type of sequence concerned.

Large-scale duplications. Whole-genome duplications have proved surprisingly common in the wake of complete sequencing efforts — doubly so because the smallest genomes have provided the best evidence for such events. Based on large-scale genome comparisons, it is now acknowledged that *Saccharomyces cerevisiae*, the first eukaryote sequenced, is an ancient polyploid[52,53]. In this case, there has been extensive gene loss since the initial duplication event. No evidence of large-scale duplications has emerged from the *S. pombe* sequence[47], but tandem and block duplications have apparently featured prominently in many yeast species[54]. Rice, which has the smallest genome among the cereals, was initially thought to be aneuploid, but analyses using improved sequence assemblies have shown it to be an ancient polyploid that has also undergone extensive segmental and individual gene duplications[55,56]. Even the tiny-genomed *A. thaliana* has turned out to be an ancient polyploid[57], prompting suggestions that all flowering plants might have polyploidy in their ancestry[58].

Decades-old (and hotly contested) hypotheses about genome duplication in early vertebrate evolution[59] have also gained support from complete sequencing efforts[55,60]. Debate over a possible duplication event that was specific to bony fishes was also recently resolved with the publication of the *T. nigroviridis* genome sequence. In this case, it was noted first that a large portion of genes on one chromosome have a duplicate copy on another chromosome and that this is true of all chromosomes, and second that nearly all matching genes present in one copy in the human genome are found in two copies in *T. nigroviridis*[33].

These important findings probably represent only the first among many novel, and in some cases quite surprising, discoveries about the large-scale evolution of genomes to emerge from complete sequencing efforts.

Testing hypotheses about genome-size evolution. In addition to providing important general insights into the mechanisms that influence nuclear DNA content, large-scale sequence data have been used in a few cases to evaluate specific hypotheses about genome-size evolution.

In one recent example, Hughes and Piontkivska[61] compared the abundances and distributions of DNA repeats in the chicken and human genome sequences and found support for the view that natural selection maintains small genome sizes in birds. This could be for genome-level reasons such as preventing the disruption of the alignment of tiny avian chromosomes by repetitive sequences[62] and/or in response to organism-level pressures for small cell size that are related to the high metabolic demands of powered flight[63]. Interestingly, most of the non-coding segments in the chicken that align with human sequences lie far from genes and often occur in clusters that seem to be under selection for some currently unknown function(s)[35]. This raises the intriguing possibility that the dominant pressure in the evolution of avian genome size is for reduction, but that this shrinkage is halted at a certain threshold by functional constraints. This represents a reversal of many earlier theories of genome-size evolution, under which non-coding DNA was thought to accumulate until it became too costly to the host.

Size from sequence: some cautions. As important as the previously mentioned insights have been, some warnings are necessary about the exclusive use of sequence data to study genome size. First, it should be borne in mind that genome sequences are rarely 'complete', making genome sequencing not only an excessively costly but also an inaccurate means by which to assess genome size even in small-genomed eukaryotes. With *D. melanogaster*, only about two-thirds of the genome was sequenced, and the total reported size was estimated from the physical lengths of photographs of the unsequenced chromosome segments[64]. In the case of *A. thaliana*, the estimate provided by the sequencing consortium[65] proved to be 25% too low, as revealed by careful genome-size analyses using best-practice techniques[66].

Second, because technological and financial limitations currently make it necessary to choose subjects with small genomes, there is a severe inherent bias in the available genome-sequence data set. As such, one must be cautious about extrapolating findings that are based exclusively on sequenced genomes to the C-value enigma in general, especially when this combines information from prokaryotes and eukaryotes (which show fundamental differences in genome organization). For example, gene number and genome size are occasionally plotted for prokaryotes and sequenced eukaryotes taken together, which gives the impression that the two parameters are strongly correlated (although they taper off weakly at higher values) across both types of organism[28,67]. However, such analyses based on data sets that are strongly biased towards small genomes do not properly convey the sharp tailing off of the relationship that is seen when the full diversity of eukaryotic genome size is considered (FIG. 2).

Third, models of genome-size evolution that are developed using sequence information alone are potentially misleading because they are necessarily based on only a small number of data. This difficulty applies equally to comparisons of a small number of complete genomes or of only a few types of sequence. By way of example, the model of Lynch and Conery[28] was based on a small number of sequenced prokaryotes and eukaryotes and proposed that much variation in

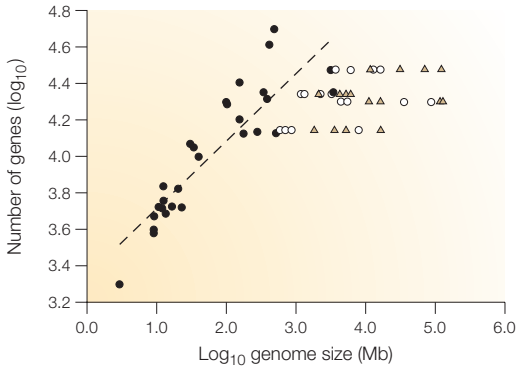


Figure 2 | **Using only sequenced genomes in genome-size studies can be misleading.** Owing to technical and fiscal constraints, the data set of sequenced genomes consists almost entirely of species with small genomes, and therefore omits most of the variation in genome size that is found in eukaryotes. This can distort the view of how genome size relates to other features. For example, comparing only species for which the genomes have been sequenced (black circles) gives the impression of a strong positive relationship (dashed line) between genome size and gene number in eukaryotes (prokaryotes are not included here, but their addition strongly reinforces this impression). However, merely plotting the mean (white circles) or maximum (yellow triangles) genome sizes (let alone the full spread of data) for the principal groups of animals and plants, and assuming that gene estimates from sequenced genomes are typical of their respective taxa, makes it clear that the overall relationship tails off sharply, beginning at a small genome size.

genome size results from differences in population size; however, one of the explicit predictions of this model (that is, that carnivores, which have smaller population sizes, should have larger genomes than herbivores) was not supported when it was tested through an analysis of the actual mammalian genome-size data[68]. Similarly, models under which genome size is shaped primarily by variation in small insertion–deletion biases[69] might have been unduly influenced by the small number of species and sequence types that have been analysed[50].

The relevance of size for sequencing

Most of the discussion so far has emphasized the light being shed on the old C-value enigma by the new field of complete genome sequencing, but there are also several ways that genome size can be of use to the study of genome sequences. The most obvious is that genome size directly influences the cost and difficulty of sequencing projects, and is therefore a primary consideration in choosing future sequencing subjects[70,71]. In fact, genome-size information is now considered a prerequisite by many of the agencies that provide funding for large-scale sequencing initiatives. Genome-size variation can also influence related molecular techniques in addition to sequencing, including the construction of genomic libraries and the amplification of specific genomic fragments by PCR[72,73].

However, knowledge of genome size is more than a practical necessity in large-scale sequencing programmes; it can also be of use in understanding other key features including structure, organization and composition. For example, genome size provides an instant approximation of the amount of non-coding DNA present in a given genome (FIG. 1), and can set the context for comparisons of constituent sequences and their relative abundances, the configurations of chromosomes (for example, in terms of heterochromatin content and distribution), and the higher-order phenotypic consequences of transposable elements and other non-coding sequences. An appreciation for the broad importance of genome-size study should also encourage the development of a set of parameters to be reported as a matter of course in whole-scale sequencing reports, such as a breakdown and total abundance estimate of transposable elements in each sequenced genome. So far there is little standardization in this regard (and many sequencing reports do not provide this information at all), which stifles efforts at broader comparisons of this key genomic feature.

For the above reasons it is not uncommon for genome-size researchers to be overwhelmed with requests to carry out new estimates on organisms of particular interest to genomics and other biological disciplines. Unfortunately there is limited networking among genome-size researchers and little dedicated financial support for such genome-sizing services, which can present a significant impediment to sequencers who are unable to locate an available collaborator. In some cases sequencing groups have made a commendable attempt to obtain original genome-size estimates themselves[33], but they did not make use of established best-practice techniques. More positively, recent advances in genome-size methodology and a growing worldwide interest in genome-size research indicate that only a comparatively small investment would be needed to create an effective network of genome-size specialists. This would not only become a crucial resource for other large-scale efforts in genomics such as complete genome sequencing, but would also greatly accelerate the pace of discovery in basic genome-size research.

Genome-size databases: a partially tapped resource. The inclusion of new C-value estimates — or, far less desirably, presenting only partial (for example, euchromatic) genome sizes — is not always necessary for genome-sequencing projects, because comprehensive online databases of published genome-size data have been available for several years for both plants (since 1997) and animals (since 2001). At the time of this writing, the Plant DNA C-values Database contains data for about 4,840 species, and the Animal Genome Size Database covers another 4,060 species. Although members of the genome-sequencing community are frequent visitors to these databases, any mention of previously published genome-size data has remained curiously absent from many complete sequence reports. For example, when the genome sequence of *A. thaliana*

was described[65], 10 individual estimates of its genome size had already been published and were available in the online database, but none was cited[66].

It is possible that many genome sequencers remain unaware of these databases, are unfamiliar with the units of genome-size used (picograms, where 1 pg = 10^{-12} g = 978 Mb), or are put off by the disagreement found between multiple entries for some species. Inevitably, the inclusion of all estimates made over the past 50 years does introduce a degree of error into the databases[74], making quality control an important issue (as is also true for other repositories of genetic information[75,76]). Fortunately, the recent development of more accurate methods and standardized protocols promise to improve the consistency of the data set in the future[77–80]. These issues aside, it is clear that the genome-size databases remain a partially untapped resource that can and should be used to a much greater degree by the broader community of both genome sequencers and genome-size researchers.

Synergy in experimental genomics. Experimental manipulations of non-coding DNA represent an intriguing area of overlap between sequence-based and size-based genome research. In a recent study, Nóbrega and co-authors[81] deleted two megabase-sized non-coding intervals of the mouse genome, and reported that this had no observable effects on the phenotype (at least under laboratory conditions). The emphasis in this case was more on the consequences on gene expression, as the deleted fragments represented less than 0.1% of the total mouse genome. Genome-size researchers have also recognized the potential utility of experimental manipulations that involve either deletions or injections of DNA in directly assessing the impacts of DNA content on cell sizes and division rates[13]. This would probably involve large changes in DNA content carried out in cell culture, but the principle is similar and the results of both types of study would clearly be mutually informative. In combination with continued cross-species comparisons, experimental work such as this could shed light on how the abundance, type and location of non-coding DNA affects parameters that range from gene expression to cellular and organismal phenotypes.

A new genomic enigma. The strikingly low number of genes required to construct even the most complex organisms represents one of the most surprising findings to emerge from the analysis of complete genome sequences. Whereas previous estimates of the human gene number had ranged from 60,000–120,000, the draft sequence indicated a mere 30,000–35,000 (REF. 16), a total that was further reduced to only 20,000–25,000 in the final assembly[82]. Almost immediately, this was offered as a new 'G-value paradox' or 'N-value paradox', in direct reference to the previous C-value paradox[83–85]. The similarity of the two paradoxes indicates that some conceptual lessons from the past 5 decades of genome-size study could be of use in the next phase of genome-sequence research.

The analogy with the C-value paradox is apt for three reasons. First, because the reason for the surprise was similar in both cases: namely, inappropriate assumptions about how simple quantitative aspects of the genome should determine organism-level complexity. Second, because in both cases the initial paradox was solved by a simple realization: with the C-value paradox that not all (or even much) eukaryotic DNA codes for proteins, and with the more recent G-value paradox that individual genes can code for multiple products, and that regulation and expression are more important than number. And third, because the solution to both paradoxes touched off a complex series of puzzles that will keep genome biologists busy for many years to come. Therefore, the new G-value enigma includes many questions that relate to the frequencies, mechanistic bases, and impacts of processes such as shifts in regulatory pathways, chromosomal rearrangements, alternative splicing, and gene–gene, gene–protein and protein–protein interactions. The crucial point, as with the C-value enigma, will be to overcome the temptation to seek simple, one-dimensional explanations to this non-paradoxical puzzle.

The future of sequence and size

The next decade of genome research will be among the most exciting since the earliest days of the science. The continuing explosion in complete genome-sequencing projects will allow previously inaccessible aspects of the C-value enigma to be investigated directly, and increased knowledge about genome size will provide a wide-ranging empirical and conceptual context for understanding the large-scale evolution of eukaryotic genomes. The study of specific questions — such as the frequency and consequences of duplications (ranging in scale from individual nucleotides to entire genomes) and the locations and possible structural or regulatory roles of existing and extinct transposable elements — will be enlightening to both subdisciplines. This work will also have substantial impacts on evolutionary theory, given the crucial roles that these 'non-standard' genetic processes have had in major evolutionary transitions[22].

Long gone are the days when genomes could be considered to be strings of independently functioning genes, each coding for a single protein product and interrupted by lengthy but irrelevant stretches of non-coding DNA. The evolution of both the genic and non-coding portions of genomes has proved to be a complex issue that requires, first and foremost, an appreciation of genomes as integrated levels of biological organization with their own inherent evolutionary processes and histories. Deciphering how genomes come to acquire their characteristics, and how these in turn affect the evolution of features at higher levels of organization, will demand a broadly integrative approach that synergizes insights from various disciplines. Closing the gap between sequence and size will mark a powerful first step in this challenging but exciting enterprise.

1. Boivin, A., Vendrely, R. & Vendrely, C. L'acide désoxyribonucléique du noyau cellulaire dépositaire des caractères héréditaires; arguments d'ordre analytique. *C. R. Acad. Sci.* **226**, 1061–1063 (1948) (in French).
2. Mirsky, A. E. & Ris, H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* **34**, 451–462 (1951).
3. Thomas, C. A. The genetic organization of chromosomes. *Annu. Rev. Genet.* **5**, 237–256 (1971).
4. Ohno, S. in *Evolution of Genetic Systems* (ed. Smith, H. H.) 366–370 (Gordon and Breach, New York, 1972).
5. Comings, D. E. The structure and function of chromatin. *Adv. Hum. Genet.* **3**, 237–431 (1972).
6. Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* **34**, 247–278 (1978).
7. Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
8. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
9. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
10. Gregory, T. R. in *The Evolution of the Genome* (ed. Gregory, T. R.) 3–87 (Elsevier, San Diego, 2005).
The author provides a comprehensive recent review of the evolution of genome size in animals.
11. Bennett, M. D. & Leitch, I. J. in *The Evolution of the Genome* (ed. Gregory, T. R.) 89–162 (Elsevier, San Diego, 2005).
The authors provide a comprehensive recent review of the evolution of genome size in plants.
12. Filipski, A. & Kumar, S. in *The Evolution of the Genome* (ed. Gregory, T. R.) 521–583 (Elsevier, San Diego, 2005).
13. Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev.* **76**, 65–101 (2001).
This article outlines the key concepts in the study of the C-value enigma and the main theories that have been proposed to explain it.
14. Gregory, T. R. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. *Blood Cells Mol. Dis.* **27**, 830–843 (2001).
15. Gregory, T. R. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* **95**, 133–146 (2005).
16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
This paper provides the first whole-scale view of the contents and characteristics of a relatively large animal genome.
17. Kidwell, M. G. in *The Evolution of the Genome* (ed. Gregory, T. R.) 165–221 (Elsevier, San Diego, 2005).
18. Kidwell, M. G. & Lisch, D. R. Transposable elements and host genome evolution. *Trends Ecol. Evol.* **15**, 95–99 (2000).
19. Kidwell, M. G. & Lisch, D. R. Transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**, 1–24 (2001).
20. Brookfield, J. F. Y. The ecology of the genome — mobile DNA elements and their hosts. *Nature Rev. Genet.* **6**, 128–136 (2005).
21. Furano, A. V., Duvernell, D. D. & Boissinot, S. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* **20**, 9–14 (2004).
22. Gregory, T. R. in *The Evolution of the Genome* (ed. Gregory, T. R.) 679–729 (Elsevier, San Diego, 2005).
This chapter places emerging knowledge of genome evolution in the context of an expanded evolutionary theory, and highlights some key 'non-standard' genetic processes that have been important in various major evolutionary transitions.
23. Brookfield, J. F. Y. Mobile DNAs: the poacher turned gamekeeper. *Curr. Biol.* **13**, R846–R847 (2003).
24. Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19**, 68–72 (2003).
The authors suggest a significant role for formerly parasitic elements in the evolution and function of complex genomes.
25. Zhou, L. *et al.* Transposition of *hAT* elements links transposable elements and V(D)J recombination. *Nature* **432**, 995–1001 (2004).
This article provides intriguing evidence for a link between formerly parasitic genomic elements and the evolution of the adaptive immune system of vertebrates.
26. Kimura, R. H., Choudary, P. V. & Schmid, C. W. Silk worm *Bm1* SINE RNA increases following cellular insults. *Nucleic Acids Res.* **27**, 3380–3387 (1999).
27. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).
28. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
29. 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** (Suppl. A), 37–44 (1998).
This paper describes the extraordinary influence that transposable elements can have on the evolution of genome size, even over relatively short timescales.
30. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**, 933–951 (2004).
31. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
32. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
33. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957 (2004).
This article provides some of the most compelling evidence so far that a complete round of genome duplication occurred in an early ancestor of the bony fishes.
34. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).
35. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
36. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
37. Kumar, A. & Bennetzen, J. L. Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532 (1999).
38. Xia, Q. *et al.* A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**, 1937–1940 (2004).
39. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
40. Kim, J. M. *et al.* Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8**, 464–478 (1998).
41. Volff, J.-N., Lehrach, H., Reinhardt, R. & Chourrout, D. Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. *Mol. Biol. Evol.* **21**, 2022–2033 (2004).
42. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
43. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
44. Dietrich, F. S. *et al.* The *Ashbya gosypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**, 304–307 (2004).
45. Galagan, J. E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868 (2003).
46. Harrison, P. M. *et al.* Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280 (2002).
47. Wood, V., Gwilliam, R. & Rajandream, M.-A. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
48. Zbodnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
49. Petrov, D. A. Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17**, 23–28 (2001).
50. Gregory, T. R. Insertion–deletion biases and the evolution of genome size. *Gene* **324**, 15–34 (2004).
51. Zhang, Z., Carriero, N. & Gerstein, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **20**, 62–67 (2004).
52. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
53. Ochman, H., Daubin, V. & Lerat, E. A bunch of fun-guys: the whole-genome view of yeast evolution. *Trends Genet.* **21**, 1–3 (2005).
54. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).
55. Van de Peer, Y. & Meyer, A. in *The Evolution of the Genome* (ed. Gregory, T. R.) 329–368 (Elsevier, San Diego, 2005).
56. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, 267–281 (2005).
57. Simillion, C., Vanepoele, K., Van Montagu, M. C. E., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99**, 13627–13632 (2002).
The authors dicuss the ancient genome duplication that occurred in this tiny-genomed flowering plant; this evidence raises the possibility that all angiosperms have polyploidy in their ancestry.
58. Tate, J. A., Soltis, D. E. & Soltis, P. S. in *The Evolution of the Genome* (ed. Gregory, T. R.) 371–426 (Elsevier, San Diego, 2005).
59. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
60. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31**, 200–204 (2002).
61. Hughes, A. L. & Piontkivska, H. DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. *BMC Evol. Biol.* **5**, 12 (2005).
62. Burt, D. W. Origin and evolution of avian minichromosomes. *Cytogenet. Genome Res.* **96**, 97–112 (2002).
63. Gregory, T. R. 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 (2002).
64. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
65. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
66. Bennett, M. D., Leitch, I. J., Price, H. J. & Johnston, J. S. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25 % larger than the Arabidopsis Genome Initiative estimate of ~125 Mb. *Ann. Bot.* **91**, 547–557 (2003).
This paper demonstrates the crucial importance of using best-practice techniques in the analysis of genome size, and highlights the potential problems involved in estimating genome size by using only sequence data.
67. Hughes, A. L. *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, Oxford, UK, 1999).
68. Vinogradov, A. E. Testing genome complexity. *Science* **304**, 389–390 (2004).
69. Petrov, D. A. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61**, 533–546 (2002).
70. Pryer, K. M., Schneider, H., Zimmer, E. A. & Banks, J. A. Deciding among green plants for whole genome studies. *Trends Plant Sci.* **7**, 550–554 (2002).
71. Evans, J. D. & Gundersen-Rindal, D. Beenomes to *Bombyx*: future directions in applied insect genomics. *Genome Biol.* **4**, 107 (2003).
72. Garner, T. W. J. Genome size and microsatellites: the effect of nuclear size on amplification potential. *Genome* **45**, 212–215 (2002).
73. Fay, M. F., Cowan, R. S. & Leitch, I. J. The effects of DNA content (C-value) on the quality and utility of AFLP fingerprints. *Ann. Bot.* **95**, 237–246 (2005).
74. Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann. Bot.* **95**, 45–90 (2005).
75. Pennisi, E. Keeping genome databases clean and up to date. *Science* **286**, 447–450 (1999).
76. Hadley, C. Righting the wrongs. *EMBO Rep.* **4**, 829–831 (2003).
77. Vilhar, B., Greilhuber, J., Koce, J. D., Temsch, E. M. & Dermastia, M. Plant genome size measurement with DNA image cytometry. *Ann. Bot.* **87**, 719–728 (2001).
78. Hardie, D. C., Gregory, T. R. & Hebert, P. D. N. From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. *J. Histochem. Cytochem.* **50**, 735–749 (2002).
79. DeSalle, R., Gregory, T. R. & Johnston, J. S. Preparation of samples for comparative studies of arthropod chromosomes: visualization, *in situ* hybridization, and genome size estimation. *Meth. Enzymol.* **395**, 460–488 (2005).
80. Dolezel, J. & Bartos, J. Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* **95**, 99–110 (2005).

81. Nóbrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E. M. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988–993 (2004).
82. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
83. Claverie, J.-M. What if there are only 30,000 human genes? *Science* **291**, 1255–1257 (2001).
84. Betrán, E. & Long, M. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**, 65–80 (2002).
85. Hahn, M. W. & Wray, G. A. The G-value paradox. *Evol. Dev.* **4**, 73–75 (2002).
86. Gregory, T. R. & DeSalle, R. in *The Evolution of the Genome* (ed. Gregory, T. R.) 585–675 (Elsevier, San Diego, 2005).
87. Sparrow, A. H., Price, H. J. & Underbink, A. G. in *Evolution of Genetic Systems* (ed. Smith, H. H.) 451–494 (Gordon and Breach, New York, 1972).
88. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079 (2002).
89. Bennetzen, J. L., Ma, J. & Devos, K. M. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95**, 127–132 (2005).
90. Orel, N. & Puchta, H. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. *Plant Mol. Biol.* **51**, 523–531 (2003).
91. Filkowski, J., Kovalchuk, O. & Kovalchuk, I. Dissimilar mutation and recombination rates in *Arabidopsis* and tobacco. *Plant Sci.* **166**, 265–272 (2004).
92. Taylor, J. S. & Raes, J. in *The Evolution of the Genome* (ed. Gregory, T. R.) 289–327 (Elsevier, San Diego, 2005).
93. Ohta, T. Population genetics of selfish DNA. *Nature* **292**, 648–649 (1981).
94. Hickey, D. A. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101**, 519–531 (1982).
95. Charlesworth, B. The population biology of transposable elements. *Trends Ecol. Evol.* **2**, 21–23 (1987).
96. Charlesworth, B., Sniegowski, P. & Stephan, W. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215–220 (1994).
97. Promislow, D. E. L., Jordan, I. K. & McDonald, J. F. Genomic demography: a life-history analysis of transposable element evolution. *Proc. R. Soc. Lond. B* **266**, 1555–1560 (1999).
98. Arkhipova, I. & Meselson, M. Transposable elements in sexual and ancient asexual taxa. *Proc. Natl Acad. Sci. USA* **97**, 14473–14477 (2000).
99. Hatcher, M. J. Persistence of selfish genetic elements: population structure and conflict. *Trends Ecol. Evol.* **15**, 271–277 (2000).
100. Schön, I. & Martens, K. Transposable elements and asexual reproduction. *Trends. Evol.* **15**, 287–288 (2000).
101. Wendel, J. F., Cronn, R. C., Johnston, J. S. & Price, H. J. Feast and famine in plant genomes. *Genetica* **115**, 37–47 (2002).
The authors show that genome sizes can change both by increasing and decreasing, even within a narrow taxonomic range.
102. Leitch, I. J., Soltis, D. E., Soltis, P. S. & Bennett, M. D. Evolution of DNA amounts across land plants (Embryophyta). *Ann. Bot.* **95**, 207–217 (2005).
103. Thomson, K. S. An attempt to reconstruct evolutionary changes in the cellular DNA content of lungfish. *J. Exp. Zool.* **180**, 363–372 (1972).
104. Thomson, K. S. & Muraszko, K. Estimation of cell size and DNA content in fossil fishes and amphibians. *J. Exp. Zool.* **205**, 315–320 (1978).
105. Conway Morris, S. & Harper, E. Genome size in conodonts (Chordata): inferred variations during 270 million years. *Science* **241**, 1230–1232 (1988).
106. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in a majority of angiosperms. *Science* **264**, 421–424 (1994).
107. Wong, G. K.-S., Passey, D. A., Huang, Y.-Z., Yang, Z. & Yu, J. Is 'junk' DNA mostly intron DNA? *Genome Res.* **10**, 1672–1678 (2000).
108. Wong, G. K.-S., Passey, D. A. & Yu, J. Most of the human genome is transcribed. *Genome Res.* **11**, 1975–1977 (2001).
109. Vinogradov, A. E. Intron–genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49**, 376–384 (1999).
110. Moriyama, E. N., Petrov, D. A. & Hartl, D. L. Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15**, 770–773 (1998).
111. Wendel, J. F. *et al.* Intron size and genome size in plants. *Mol. Biol. Evol.* **19**, 2346–2352 (2002).

Acknowledgments

Sincere thanks to S. Adamowicz and two of the anonymous reviewers for providing constructive criticism on an early draft of the paper.

Competing interests statement

The author declares no competing financial interests.

Online links

FURTHER INFORMATION

Animal Genome Size Database: http://www.genomesize.com
Fungal C-values Database: http://www.zbi.ee/fungal-genomesize/index.php
GOLD — Genomes OnLine Database: http://www.genomesonline.org
Plant DNA C-values Database: http://www.rbgkew.org.uk/cval/homepage.html
The Gregory Laboratory: http://www.uoguelph.ca/~rgregory
TIGR Comprehensive Microbial Resource: http://pathema.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
Access to this interactive links box is free online.