What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content

Tyler A. Elliott and T. Ryan Gregory

Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G 2W1

Some notable exceptions aside, eukaryotic genomes are distinguished from those of Bacteria and Archaea in a number of ways, including chromosome structure and number, repetitive DNA content, and the presence of introns in protein-coding regions. One of the most notable differences between eukaryotic and prokaryotic genomes is in size. Unlike their prokaryotic counterparts, eukaryotes exhibit enormous (more than 60 000-fold) variability in genome size which is not explained by differences in gene number. Genome size is known to correlate with cell size and division rate, and by extension with numerous organism-level traits such as metabolism, developmental rate or body size. Less well described are the relationships between genome size and other properties of the genome, such as gene content, transposable element content, base pair composition and related features. The rapid expansion of ‘complete’ genome sequencing projects has, for the first time, made it possible to examine these relationships across a wide range of eukaryotes in order to shed new light on the causes and correlates of genome size diversity. This study presents the results of phylogenetically informed comparisons of genome data for more than 500 species of eukaryotes. Several relationships are described between genome size and other genomic parameters, and some recommendations are presented for how these insights can be extended even more broadly in the future.

1. Introduction

A number of major characteristics distinguish the genomes of eukaryotes from those of ‘prokaryotes’ (Bacteria and Archaea). Of course, the most obvious (and defining) difference is the separation of the genome from the cytoplasm by the presence of a nuclear membrane. Other features include the packaging of the genome into multiple linear chromosomes rather than a single circular one, the division of genes into protein-coding exons and non-coding introns, and the presence of often large quantities of repetitive non-genic DNA. Exceptions to this are not uncommon: some bacteria possess linear chromosomes (e.g. Streptomyces [1]), or multiple chromosome-sized plasmids (e.g. Vibrio [2]), or mobile introns, or insertion elements. Other prokaryotes exhibit polyploidy by containing multiple copies of the genome, and some may even contain endosymbionts with their own genomes. What sets eukaryotes apart, however, is that they typically exhibit most or all of these features simultaneously.

Another key factor that sets eukaryotic genomes apart is their size. The genomes of Bacteria and Archaea are all diminutive, in the range of 140 kilobase pairs (kbp) to approximately 15 megabase pairs (Mbp), with most of this variability accounted for by differences in the number of protein-coding genes [3]. In eukaryotes, by contrast, haploid nuclear genome sizes (‘C-values’) range more than 60 000-fold, from 2.3 Mbp in the parasitic microsporidian Encephalitozoon intestinalis [4] to approximately 150 000 Mbp in the plant Paris japonica [5], with this enormous diversity bearing no relationship to any intuitive notions of organismal complexity. This is not a new observation. Perplexingly large differences in genome size among morphologically similar species and the occurrence of copulent genomes in comparatively simple organisms was first noted in the late 1940s...
and early 1950s. Twenty years later, this remained sufficiently confusing to be dubbed the ‘C-value paradox’ [6].

The discovery that eukaryotic genomes contain vast quantities of non-protein-coding DNA resolved the ‘paradox’ of a general lack of correspondence between genome size and organismal complexity. At the same time, it has raised a number of important new questions that persist as major subjects of investigation. What kinds of sequences are present in eukaryotic genomes, and which of these contribute to the extensive variability in total genome size? How do these elements accumulate or become lost in genomes over evolutionary time scales? Does this non-coding majority have any effects (positive or negative) on organismal biology, or even functions for which it has been subject to natural selection at the organismal level (sensu [7])? Why do some genomes remain (or become) streamlined, whereas others reach staggering sizes? Collectively, these questions have been considered part of a complex puzzle known as the ‘C-value enigma’ [8].

Much progress has been made in answering the various component questions of the C-value enigma over the past few decades. The implementation of rapid and reliable methods for estimating genome size has led to the accumulation of genome size estimates for more than 15,000 species of animals, plants, fungi and ‘protists’ [9–13]. Although this represents a tiny fraction of overall eukaryotic diversity (especially among hyperdiverse but badly under-represented taxa such as invertebrate animals and ‘protists’), it has highlighted a number of important patterns and correlates of genome size diversity. In the broadest terms, genome size correlates positively with cell size across a wide array of taxa and cell types (e.g. [14,15]). C-values have also been found to correlate inversely with cell division rate, meaning that large genomes are typically found in the nuclei of large, slowly dividing cells. How these cell-level correlations manifest as impacts at the organismal level varies according to the biology of the group in question, but there are numerous examples from both animals and plants of relationships between genome size and parameters such as body size, metabolism, developmental rate and geographical distribution (e.g. [14,15]).

Surprisingly, less progress has been made in understanding relationships between genome size and content, despite a rapidly expanding dataset of euchromatic genome sequences. In part, this is due to a general lack of sequences from truly large (let alone enormous) genomes—an omission borne of current technological and computational limitations in sequencing very large, repetitive genomes. Nevertheless, genome sequencing data are now available for hundreds of eukaryotic genomes, ranging in size from 2.3 Mbp in E. intestinalis to approximately 20,000 Mbp in the Norway spruce Picea abies [4,16]. Additionally, insights from large genomes are beginning to flow from survey sequencing using next-generation technologies. This study makes use of available data on genome size and sequence in order to assess potential relationships between genome size, content and organization across a diverse array of eukaryotes.

2. Material and methods

(a) Genome sequence information

Genome sequence information was compiled manually from the literature for completed, draft and survey-sequenced genomes, as were available up to September 2014. In total, 22 parameters were derived from these data, covering various metrics related to number of genes, exon and intron content, repetitive content, transposable element (TE) content, base pair composition and chromosome number. Where necessary, information from the original genome publications was supplemented using a variety of published and online sources. Details of the sources are provided along with the complete dataset in electronic supplementary material, S1. In total, the compiled dataset included information for 502 species, including 148 species of animals, 81 land plants, 202 fungi and 70 ‘protists’.

(b) Genome size data

Genome size data for these species came from two major sources: the genome sequence publications and online databases including the Animal Genome Size Database [11], the Plant DNA C-Values Database [10] and the Fungal Genome Size Database [12]. These data represent estimates of genome size obtained primarily through Feulgen densitometry (more recently, using image analysis) or flow cytometry.

(c) Statistical analyses

Summary statistics and correlation coefficients were calculated using standard methods. Values for both genome size and tabulated genome statistics were log-transformed prior to analysis. Because shared ancestry violates the assumption of independence of species data, correlations were phylogenetically corrected using Felsenstein’s [17] phylogenetically independent contrasts (PICs), positivized and forced through the origin, using the PDAP module [18] in Mesquite v. 2.75 [19]. Given the broad phylogenetic coverage of the current dataset, it was necessary to assemble phylogenetic trees manually as no single species-level tree yet exists for all eukaryotes. This was achieved using information provided in the Tree of Life Database [20] as a guide, which was supplemented with phylogenetic data from the literature when needed. Because the tree included only topology and no branch length data, branch lengths were all set to unity for PIC analyses. The resolution of the phylogeny was imperfect, meaning that there were many soft polytomes (which explains the much lower number of contrasts as compared with the number of species in the dataset). One degree of freedom was subtracted for each instance of a soft polytomy [21]. The phylogenetic tree and source references used in these analyses are provided in electronic supplementary material, S2.

3. Results

(a) Genome size

The average sizes of the sequenced genomes included in this study were 1153.98 Mbp for animals, 1065.82 Mbp for land plants and 34.98 Mbp for fungi (table 1). In comparison, the average genome sizes for these groups based on estimates in their respective databases is much larger: 4176.06 Mbp for animals (based on 5635 species [11]), 6120.79 Mbp for plants (based on 8257 species [10,13]) and 65.89 Mbp for fungi (based on 1916 estimates [12]).

In total, 223 (about 43% of the dataset, nearly all of them animals and plants) of the species included in this study had genome size estimates that were generated both by sequencing assembly and one of the densitometric or flow cytometric methods. This allowed a comparison of the consistency of genome size estimates derived from sequencing projects versus traditional cytogenetic approaches. As shown in figure 1a, the two sources of genome size estimates are highly correlated ($r = 0.9527$, $p < 0.0001$, $d = 0.9527$, $p < 0.0001$).
Table 1. Summary of genome data included in this study.

<table>
<thead>
<tr>
<th>parameter</th>
<th>animals</th>
<th>land plants</th>
<th>fungi</th>
<th>protists</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>assembled genome size (Mb)</td>
<td>1153.98 ± 100.18 (n = 149)</td>
<td>1065.82 ± 176.86 (n = 83)</td>
<td>34.98 ± 1.74 (n = 218)</td>
<td>61.11 ± 9.76 (n = 70)</td>
</tr>
<tr>
<td>estimated genome size (Mb)</td>
<td>1294.71 ± 110.65 (n = 149)</td>
<td>1498.43 ± 232.52 (n = 83)</td>
<td>35.71 ± 1.92 (n = 218)</td>
<td>75.51 ± 21.49 (n = 70)</td>
</tr>
<tr>
<td>discrepancy (Mb)</td>
<td>165.107 ± 25.73 (n = 127)</td>
<td>448.83 ± 206.26 (n = 80)</td>
<td>26.38 ± 12.34 (n = 6)</td>
<td>100.82 ± 87.56 (n = 10)</td>
</tr>
<tr>
<td>average database genome size (Mb)</td>
<td>4176.06 ± 112.36</td>
<td>6120.79 ± 107.58</td>
<td>65.89 ± 5.31</td>
<td>n.a.</td>
</tr>
<tr>
<td>maximum database genome size (Mb)</td>
<td>129,907.74 (n = 5635)</td>
<td>148,852 (n = 8257)</td>
<td>5800 (n = 1916)</td>
<td>n.a.</td>
</tr>
<tr>
<td>gene content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of protein-coding genes</td>
<td>18,943 ± 451.82 (n = 139)</td>
<td>35,577 ± 1641.08 (n = 80)</td>
<td>9953 ± 315.16 (n = 202)</td>
<td>12,589 ± 1148.69 (n = 70)</td>
</tr>
<tr>
<td>total amount of coding DNA (Mb)</td>
<td>27.58 ± 1.26 (n = 90)</td>
<td>39.23 ± 1.81 (n = 64)</td>
<td>13.059 ± 0.56 (n = 97)</td>
<td>18.55 ± 2.16 (n = 49)</td>
</tr>
<tr>
<td>coding % of estimated GS</td>
<td>10.4 ± 1.12 (n = 90)</td>
<td>7.86 ± 0.87 (n = 64)</td>
<td>46.66 ± 1.62 (n = 97)</td>
<td>42.31 ± 3.23 (n = 49)</td>
</tr>
<tr>
<td>average exon length (bp)</td>
<td>218.8 ± 9.28 (n = 70)</td>
<td>256.35 ± 6.06 (n = 55)</td>
<td>498.77 ± 41.21 (n = 72)</td>
<td>600.05 ± 53.37 (n = 40)</td>
</tr>
<tr>
<td>total amount of exonic DNA per gene (bp)</td>
<td>1489.54 ± 35.65 (n = 91)</td>
<td>1159.98 ± 27.49 (n = 63)</td>
<td>1392.89 ± 24.72 (n = 89)</td>
<td>1497.27 ± 54.49 (n = 49)</td>
</tr>
<tr>
<td>average intron length (bp)</td>
<td>2172.5 ± 255.34 (n = 72)</td>
<td>430.091 ± 28.08 (n = 50)</td>
<td>133.34 ± 6.87 (n = 80)</td>
<td>204.37 ± 20.44 (n = 44)</td>
</tr>
<tr>
<td>average number of introns per gene</td>
<td>5.05 ± 0.47 (n = 26)</td>
<td>3.94 ± 0.38 (n = 20)</td>
<td>1.72 ± 0.24 (n = 50)</td>
<td>2.61 ± 0.66 (n = 27)</td>
</tr>
<tr>
<td>total amount of intronic DNA per gene (bp)</td>
<td>8191.5 ± 2033.71 (n = 25)</td>
<td>1804.45 ± 287.43 (n = 18)</td>
<td>201.23 ± 25.23 (n = 38)</td>
<td>1047.99 ± 472.46 (n = 19)</td>
</tr>
<tr>
<td>total gene region size (introns + exons) (bp)</td>
<td>9533.11 ± 2050.22 (n = 25)</td>
<td>2956.72 ± 302.28 (n = 18)</td>
<td>1655.47 ± 48.73 (n = 30)</td>
<td>2487.97 ± 552.27 (n = 17)</td>
</tr>
<tr>
<td>repetitive content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>repeats as % of assembly GS</td>
<td>27.35 ± 1.83 (n = 102)</td>
<td>50.6 ± 3 (n = 54)</td>
<td>14.38 ± 1.75 (n = 92)</td>
<td>19.45 ± 3.98 (n = 26)</td>
</tr>
<tr>
<td>total amount of repetitive DNA (Mb)</td>
<td>459.41 ± 69.96 (n = 102)</td>
<td>946.23 ± 202.51 (n = 54)</td>
<td>8.81 ± 1.76 (n = 92)</td>
<td>24.88 ± 7.56 (n = 26)</td>
</tr>
<tr>
<td>TE % of assembly GS</td>
<td>23 ± 1.85 (n = 100)</td>
<td>38.88 ± 2.44 (n = 61)</td>
<td>13.59 ± 2.75 (n = 74)</td>
<td>13.84 ± 3.19 (n = 28)</td>
</tr>
<tr>
<td>base pair composition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC %</td>
<td>37.68 ± 0.55 (n = 76)</td>
<td>36 ± 0.62 (n = 31)</td>
<td>45.73 ± 0.58 (n = 161)</td>
<td>47.24 ± 1.76 (n = 56)</td>
</tr>
</tbody>
</table>
genome size, with most of the differences reflecting a lower estimate derived from sequence assembly. (Online version in colour.)

There are discrepancies. (sequence assemblies versus estimates based on Feulgen densitometry or flow cytometry. The two sets of estimates are strongly positively correlated, although there are discrepancies according to method of estimation and genome size. The discrepancy becomes larger with increasing genome size, with most of the differences reflecting a lower estimate derived from sequence assembly. (Online version in colour.)

The great majority of these cases (more than 87%) involved a discrepancy between methods of 10% or more. The majority of cases in which assembly estimates were the larger, most were from smaller genomes, leading to an average mismatch of only about 90 Mbp in these cases. As seen in figure 1b, the relative discrepancy between methods became more pronounced as genome size increased.

(b) Gene content

Gene content of sequenced genomes was evaluated according to several metrics, which are summarized in table 1. Overall, fungi exhibited the lowest average number of protein-coding genes per genome (9953). The average animal genome contains nearly twice as many genes (18 943) as fungal genomes, and plants have almost twice as many as animals (35 577). Total gene number ranged at least 50-fold across available eukaryotes, nearly twice as many genes (18 943) as fungal genomes, and across the eukaryotes examined, the average number of protein-coding genes varied at least 50-fold across available eukaryotes, ranging from 95 000–124 000 in the hexaploid wheat Triticum aestivum.

However, because of their differences in total genome size, a significantly higher percentage of fungal genomes (approx. 47%) is made up of coding exons than in animals (approx. 10%) or plants (approx. 8%). Across the eukaryotes examined, total exonic content ranged from 1.98 Mbp (86% of the genome) in E. intestinalis to approximately 95 000–124 000 in the hexaploid wheat Triticum aestivum.

The repetitive component of the available genomes differed significantly between fungi (approx. 14% of the genome, on the three taxa, reflecting a difference in the average number of introns per gene (table 1). The average lengths of introns were much larger in animals (approx. 2200 bp) than in plants (approx. 430 bp) or fungi (approx. 130 bp). A summary of these parameters is provided for ‘protists’ in table 1, but note that this does not represent a single phylogenetic group and averages should therefore be interpreted with caution.

Across all eukaryotes, the number of protein-coding genes correlated positively with genome size ($r = 0.5288$, $p < 0.0001$, $n = 490$ contrasts; figure 2a), as did the absolute amount of DNA consisting of exons ($r = 0.5275$, $p < 0.0001$, $n = 299$ contrasts). Positive correlations were also found between genome size and both gene number and coding region percentage within animals, plants and fungi when these groups were analysed individually (all $p < 0.05$; figure 2b). Across all eukaryotes, genome size was correlated with average exon length ($r = 0.2468$, $p < 0.001$, $n = 236$ contrasts), but not with the total amount of exonic DNA per gene ($r = 0.0996$, $p = 0.5288$, $n = 291$ contrasts) and the total amount of exonic DNA per gene ($r = 0.0996$, $p = 0.5288$, $n = 291$ contrasts). The percentage of the genome consisting of protein-coding exons was inversely correlated with genome size within and across eukaryotes ($r = -0.6302$, $p < 0.0001$, $n = 300$ contrasts).

Intron content and genome size were positively correlated across eukaryotes in terms of average intron size ($r = 0.6065$, $p < 0.0001$, $n = 245$ contrasts), the number of introns per genome ($r = 0.4535$, $p < 10^{-7}$, $n = 121$ contrasts) and the total amount of intronic DNA present in the genome ($r = 0.6079$, $p < 10^{-11}$, $n = 115$ contrasts; figure 2c).

(c) Repetitive content versus genome size

The repetitive component of the available genomes differed significantly between fungi (approx. 14% of the genome, on
average), animals (approx. 27%) and plants (approx. 51%). The majority (more than 75%) of this repetitive content was identified as TEs (table 1). The proportion of the genome composed of repetitive elements was positively correlated with genome size across eukaryotes ($r = 0.4296$, $p = 0.0001$, $n = 273$ contrasts; figure 3a) as well as within animals ($r = 0.3773$, $p = 1.04 \times 10^{-2}$, $n = 101$ contrasts), plants ($r = 0.5948$, $p = 1.93 \times 10^{-2}$, $n = 53$ contrasts) and fungi ($r = 0.5267$, $p = 3.03 \times 10^{-2}$, $n = 73$ contrasts). Likewise, the proportion of the genome consisting of TEs was positively related to genome size in the combined dataset ($r = 0.4034$, $p < 0.0001$, $n = 260$ contrasts; figure 3b) and within animals ($r = 0.53$, $p = 2.27 \times 10^{-8}$, $n = 99$ contrasts), plants ($r = 0.3679$, $p = 0.003$, $n = 60$ contrasts) and fungi ($r = 0.5267$, $p = 3.03 \times 10^{-8}$, $n = 73$ contrasts).

Because many TEs may reside within introns, a phylogenetically corrected correlation was run between the residuals of TE proportion versus genome size and intron proportion versus genome size. These two set of residuals were not significantly correlated ($r = 0.01919$, $p = 0.884$, $n = 60$ contrasts), suggesting that much of the relationship with intron content could be encompassed within the relationship with TE content.

(d) Chromosome number and base pair composition
Chromosome number in the current dataset ranged from $n = 3$ in various animals to $n = 84$ in the sea lamprey *Petromyzon marinus*. There was a weak but significant positive correlation between genome size and chromosome number across all eukaryotes ($r = 0.1456$, $p = 0.0076$, $n = 334$ contrasts).

Base pair composition (given as overall % GC) was not found to correlate significantly with genome size across all available eukaryote data ($r = -0.0754$, $p = 0.176$, $n = 323$ contrasts; figure 4), nor within animals ($r = 0.2168$, $p = 0.143$, $n = 75$ contrasts). However, % GC content was significantly negatively correlated with genome size within plants ($r = -0.4527$, $p = 0.008$, $n = 31$) and fungi ($r = -0.33$, $p = 1.26 \times 10^{-2}$, $n = 160$ contrasts).

4. Discussion
(a) Sequence and size
The rate of growth in available genome sequence information has been staggering. From the completion of the draft human genome less than 15 years ago, the global dataset has
expanded to include hundreds of species from across the
 eukaryote phylogeny. This still represents only a very tiny
fraction of total eukaryotic diversity, but it has made it
possible for the first time to examine genome content at a
truly broad scale. It has also made it possible to address
some of the longest standing questions in genome biology
in new ways. This includes an ability to assess the composi-
tional differences of genomes that range in size—a key
part of any attempt to resolve the C-value enigma.

The first notable finding that emerges in a comparison of
genome sequence information and existing genome size esti-
mates is that they are highly compatible. Indeed, there is a
very strong correspondence between genome size estimates
derived from genome assembly and those generated using
more traditional cytogenetic approaches such as Feulgen
densitometry and flow cytometry (figure 1a). There are discrep-
ancies, however. For the most part, sequence-based
calculations provide a lower estimate of total DNA content
than densitometric or flow cytometric methods. This mismatch
becomes increasingly pronounced as genome size increases
(figure 1b). Several possible explanations exist for this. The
first is that the sizes of larger genomes are systematically over-
estimated using cytogenetic methods. However, this seems
unlikely given what is known about the sources of error in
those methods and the tendency for large genomes to be under-
estimated [22]. An opposing interpretation is that sequence
data sets for larger genomes are less complete than for smaller
genomes, due to the inherent difficulties in sequencing and
assembly of highly repetitive genomes (e.g. [23]). The presence
of more non-coding DNA, much of which will be tightly com-
pacted as heterochromatin, can also make it challenging to
achieve truly complete coverage of a genome sequence. Thus,
there may be a trend towards systematic underestimation of
genome content in most sequencing projects [10]. It is worth
noting that the relationship between total coverage of a
genome and the size of the discrepancy was weakly negative
but marginally non-significant ($r = -0.117, p = 0.088, n = 212$ contrasts), suggesting that at least some of the information

---

**Figure 3.** Relationships between genome size and (a) the proportion of the genome made up of repetitive sequences in general, and (b) transposable elements in particular. (Online version in colour.)

**Figure 4.** Genome size and base pair composition (given as % GC). (Online version in colour.)
that may be missing from many sequence datasets is indeed difficult to obtain. As an example, Koga [24] showed that harder to assemble internal regions of the Tol2 DNA transposon in the *Oryzias latipes* genome were missing from the assembly but were nonetheless quantifiable using a Southern blot analysis of genomic DNA.

Perhaps more important is the disparity in the sizes of genomes that have been examined using sequencing versus traditional methods. Whereas the average for sequenced genomes is around 1000 Mbp for animals and plants, the mean values in the databases (based on thousands of species) is more than 4000 Mbp for animals and more than 6000 Mbp for plants. And this is to say nothing of the maximum reported genome sizes in animals and plants, which are 100× larger than the average sequenced genomes (table 1). The current comparative analysis, while very informative, must therefore be interpreted with the appropriate caveats. Unfortunately, it is likely to be some time before a comparative dataset of ‘complete’ genome sequences becomes available for exceptionally large genomes, due to both the cost (though less so all the time) and the technical and computational difficulty involved [25].

(b) Genome size versus gene content

The original ‘C-value paradox’ was based on the perplexing observation that total genome size was unrelated to organism complexity, which in turn was taken as a plausible indicator of the number of genes in a genome. As noted, the discovery of diverse types and large quantities of non-coding DNA alleviated this paradox decades ago. More recently, it has become clear in the post-genomic era that gene number itself is, at best, very weakly linked to complexity. Comparisons of gene numbers among early sequencing targets (e.g. humans, flies and nematodes) indicated that intuitively complex organisms may possess a surprisingly low number of genes and, conversely, that seemingly simple organisms may have higher than expected numbers of genes. Indeed, it was even suggested that there is a ‘G-value paradox’ or ‘N-value paradox’ in reference to this disconnect between gene number and complexity [26,27].

To complicate matters further, the present analysis indicates that, in fact, genome size and gene number are positively correlated across a broad array of eukaryotes (figure 2a). Some previous analyses had suggested that this may be the case (e.g. [28,29]), but these were based on a much smaller and more taxonomically limited sample of genomes and did not take phylogeny into account. In addition, an early comparative analysis of animals and plants showed that ribosomal gene (rDNA) copy number correlates positively with genome size [30]. On the other hand, the percentage of the genome composed of protein-coding sequences decreases as genomes grow larger (figure 2b), and overall it is clear that protein-coding regions contribute only a very small amount to overall genome size in all but the smallest eukaryotic genomes. This differs markedly from the situation in prokaryotes, in which protein-coding gene number is the primary determinant of genome size diversity (e.g. [31]).

There are several (non-mutually exclusive) explanations that can be offered to account for the relationship between genome size and gene number in eukaryotes. The first, and simplest, is that this reflects a legacy of relatively common (ancient) genome duplications among eukaryotic lineages. A much higher rate of polyploidy may also explain the fact that plants have nearly twice as many genes as animals on average (table 1). It may also be the case that the frequency and/or magnitude of smaller scale gene duplication (and/or loss) events scale with genome size, for example as a result of unequal crossing over between non-homologous repetitive elements. It is also possible that the retention of gene duplicates correlates with genome size, perhaps indirectly. For example, Lynch & Conery [32] argue that larger genomes evolve in lineages with smaller long-term effective population size because this allows mildly deleterious insertions of non-coding DNA to accumulate by drift rather than being eliminated by purifying selection. This same effect could lead to the accumulation of mutations leading to subfunctionalization of gene duplicates, resulting in a larger number of genes in the genomes that also increase in size through the addition of non-coding elements.

In addition to protein-coding regions, it is clear that intron content increases along with genome size (figure 2c). A positive relationship between intron content and genome size had been suggested to exist in several previous studies that examined a much smaller number of species (e.g. [33–36]), but it can now be confirmed to apply across eukaryotes in general. Introns alone are not likely to explain more than a relatively small portion of genome size diversity, but they represent another major type of genomic element that scales with total genome size, along with protein-coding regions and rDNA genes.

(c) Repetitive DNA

It has been suggested many times that repetitive DNA, and in particular TEs, is the dominant contributor to eukaryotic genome size (e.g. [15,37–40]). The results of this study confirm that repeat content in general and TE content in particular are strongly positively correlated with genome size (figure 3). Larger genomes contain proportionately more TEs than smaller ones, and at least 75% of identifiable repeats in the genomes examined are TEs. The enormous predominance of TEs has major implications for the understanding of eukaryotic genome structure, function and evolution. TEs can be important mutagens for good or ill, on the one hand potentially contributing genetic diversity upon which natural selection can act (e.g. [41,42]) and on the other hand causing deleterious mutations associated with a number of diseases (e.g. [43,44]). Moreover, the list of examples in which individual TEs have been co-opted into important regulatory, coding or structural roles in the genome continues to grow (e.g. [45]). Mechanisms that are thought to have evolved initially as defences against TE activity have likewise been co-opted for much broader regulatory roles, thereby contributing in a major way to the subsequent diversification of multicellular eukaryote lineages [46].

Of course, TEs represent a diverse set of sequences with a range of biological attributes. Some (e.g. LINEs) are truly autonomous, capable of enabling their own transposition and replication. Others (e.g. SINEs) are reliant on the molecular machinery of other TEs. Still others (DNA elements) eschew the copy-and-paste process of retrotransposition involving an RNA intermediate and instead transpose through a direct cut-and-paste mechanism. Within the major categories (Class I, or retrotransposons and Class II, or DNA
transposons) there are multiple lineages with very different evolutionary histories and propensities. These lineages can be further classified into superfamilies and families. In many genomes (e.g. human), TEs may be very numerous but largely inactive. In others (e.g. pufferfish), each type of TE may be present in low copy numbers but remain active. Some early comparisons suggested that there may in fact be an inverse correlation between genome size and TE diversity. For example, the pufferfish genome contains far more types of TEs than the human genome, despite being only one-tenth as large [9]. However, a recent large-scale comparison by the present authors revealed that any relationship between genome size and TE diversity is much more complex [47]. The results of that analysis suggest that there is an increase in TE diversity with genome size up to a point (approx. 500 Mbp), a peak in the range and maximum of TE diversity in genomes around 500–1500 Mbp in size, and then a lower level of TE diversity in genomes larger than this. The explanation for this pattern is not yet clear, but it may indicate that larger genomes evolve primarily through the expansion of only a small subset of existing TEs, perhaps at the expense of other types of TEs that had previously occupied the genome [48].

An important question is the degree to which TE content continues to expand in truly enormous genomes. As noted, the analysis presented here is based on a limited sample of total genome size diversity because thus far genome sequencing projects have been restricted to species with comparatively small genomes. Fortunately, a number of recent studies have begun to address this question using an approach based on survey sequencing, whereby a small percentage of the genome is sampled from which extrapolations to total genome content can be made [25,48,49].

Using this approach, it has been possible to investigate the enormous (approx. 20,000–50,000 Mbp) genomes of the Australian lungfish [50], salamanders [51–53], pine [54,55] and lily [56,57]. These studies indicate that TEs do indeed represent a major portion of these exceptionally large genomes.

(d) Composition and structure
Chromosome number was weakly positively correlated with genome size across all eukaryotes, which may be consistent with an influence of ancient genome duplication. Significant relationships between GC content and genome size had previously been reported in vertebrates [58] and plants [59].

(e) Looking ahead
The present analysis has provided the first large-scale assessment of relationships between genome size and a number of other genomic properties relating to both protein-coding and non-coding regions. Many of these relationships had been suspected to exist or had been reported in much narrower comparisons, but this is the first time that they have been assessed across hundreds of eukaryotes within a phylogenetic framework. However, as informative as this has been, it is nonetheless based on a very limited sample relative to overall eukaryotic species and genome size diversity. An obvious next step would therefore be to expand these analyses to cover both greater phylogenetic breadth and a much larger range in genome sizes in animals, plants and other eukaryotes with truly large genomes. Some of this expansion can be achieved relatively easily, whereas other aspects will remain technically challenging.

It is important to note that the more than 500 species included in this study represent only a subset of all sequencing projects that have been carried out to date. Raw data exist for many additional species, but it was not possible to include them in the format of the present analysis simply because the publications reporting the genome sequencing project results do not provide the basic summary required. This issue had previously been noted as particularly problematic when it comes to reporting details of TE content (both abundance and diversity [47]). This study has revealed that a similar issue pertains to other types of genomic sequences, including coding regions and introns. This situation could be drastically improved by developing and implementing some best practices for analysing and reporting genome sequence information. In particular, a set of standardized metrics relating to basic genome composition would go a very long way towards facilitating large-scale comparisons, and thereby developing a better understanding of eukaryotic genome diversity on the broadest scale. This could include details of base pair composition, chromosome number and size, gene number, intron number and size, total repeat content, and TE abundance, diversity and activity. Other parameters not covered in this study could also be added, including the number of processed and classical pseudogenes, heterochromatin content, variable number tandem repeats (minisatellites and microsatellites), relative centromere and telomere size, and other such details of composition and structure.

Standardized reporting of genome properties would greatly improve the situation moving forward, but there remain hundreds of existing genome sequences that have yet to be characterized in this way. The development and use of a standardized bioinformatics pipeline to glean this information efficiently could help to fill this gap, and the creation of a curated, user-friendly, open-access database would make this information accessible and usable in large-scale comparative analyses. Any new sequence data could then be added to the database.

More challenging will be the expansion of the dataset to include detailed information about larger genomes. Survey sequencing provides a valuable snapshot of genomic contents, but obviously is less informative than ‘complete’ sequence data. The major challenges in sequencing large genomes stem from their highly repetitive nature and the resulting difficulties in both sequencing and subsequent assembly. Technology continues to advance rapidly, however, and there is reason for optimism that these challenges may eventually be overcome. One of the current issues is that many of the repeats of interest are longer than the individual sequence reads, but this may be ameliorated with the advent of long-read technology (e.g. [60,61]) or through the synthetic generation of longer reads (e.g. [62]). For enormous genomes like those of lungfishes, aquatic salamanders, pine and lilies, it may be necessary to sort chromosomes for individual sequencing, as has been done for some smaller genomes (e.g. [63]) or to use haploid DNA and invoke novel assembly approaches (e.g. [64]). Gaining insights into the composition of even the largest genomes is very difficult, but it is not impossible.

(f) Concluding remarks
The current analysis has highlighted some intriguing patterns with regard to genome size and content. In keeping with
expectation, genome size is positively correlated with repetitive DNA amount across eukaryotes. TEs represent the bulk of repetitive content, and it is clear that as genome size increases, a higher proportion of the genome is made up of TE sequences. However, such relationships are not restricted to repetitive elements. Intron content also scales with genome size, as does the number of protein-coding genes and the amount of coding material per genome. Combined with other suggested relationships with genome size (e.g. rDNA copy number [30]; centromere size [65]), a pattern is emerging in which most or all components of eukaryotic genomes scale up in content as genomes become larger. Testing this and other hypotheses about genome structure, composition and size will require the use of additional data, especially from species with much larger genomes than have typically been studied to date. However, given the extraordinary rapidity in the growth of available data in the past 10–15 years, there is every reason to anticipate that these questions will become answerable in the near future.

References


