Repeat sequence turnover shifts fundamentally in species with large genomes

Petr Novák[1], Maïté S. Guignard[2,3], Pavel Neumann[1], Laura J. Kelly[2,3], Jelena Mlinarec[4], Andrea Koblížková[1], Steven Dodsworth[3,5], Aleš Kovařík[6], Jaume Pellicer[2], Wencai Wang[3,8], Jiří Macas[1*], Ilia J. Leitch[2*], Andrew R. Leitch[3*]

[1] Biology Centre, Czech Academy of Sciences, České Budějovice, CZ-37005, Czech Republic

[2] Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK

[3] School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK

[4] Division of Molecular Biology, Department of Biology, University of Zagreb, Croatia

[5] School of Life Sciences, University of Bedfordshire, Luton LU1 3JU, UK

[6] Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

[7] Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Passeig del Migdia sn, 08038 Barcelona, Catalonia, Spain

[8] Guangzhou University of Chinese Medicine, Guangzhou, 510405, China

*Authors for Correspondence:

Ilia J Leitch, Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK. Telephone: +44 208 332 5329. Email: i.leitch@kew.org

Jiří Macas, Biology Centre, Czech Academy of Sciences, České Budějovice, CZ-37005, Czech Republic. Telephone: + 420 38 777 5516. Email: macas@umbr.cas.cz

Andrew R Leitch, School of Biological and Chemical Sciences, Queen Mary University of London. London E1 4NS. Telephone: +44 207 882 5294. Email: a.r.leitch@qmul.ac.uk

Petr Novák, petr@umbr.cas.cz

Ilia J. Leitch, i.leitch@kew.org

Jelena Mlinarec, jelena@biol.pmf.hr

Laura Kelly, l.kelly@kew.org

Andrea Koblížková, andrea@umbr.cas.cz

Pavel Neumann, neumann@umbr.cas.cz

Steven Dodsworth, steven.dodsworth@beds.ac.uk

Maïté Guignard, maiteguignard@gmail.com

Wencai Wang, wencaiwang@gzucm.edu.cn

Aleš Kovřík, kovarik@ibp.cz

Jaume Pellicer, J.Pellicer@kew.org

Jiří Macas, macas@umbr.cas.cz

Andrew R. Leitch, a.r.leitch@qmul.ac.uk

Abstract

Given the extraordinary 2,400-fold range of genome sizes (0.06-148.9 Gbp/1C) encountered in seed plants (angiosperms and gymnosperms) and a broadly similar gene content (amounting to ~0.03 Gbp/1C), one might predict the repeat component of the genome will increase with genome size, resulting in the largest genomes being almost entirely repetitive. We test this prediction using the same bioinformatic approach for 101 species to ensure consistency in what constitutes a repeat. We reveal a fundamental change in repeat turnover in genomes above ~ 10 Gbp/1C, such that species with the largest genomes are only about 50% repetitive. Given that genome size impacts many plant traits, habits and life strategies, this fundamental shift in repeat dynamics will likely impact the evolutionary trajectory of species lineages themselves.

Main

There is an increasing realisation of the importance of repeat sequences in the activity, functioning and evolution of the genome [1,2]. There are also increasing amounts of data on the repeat composition of different plant species, including from whole genome sequences for over 300 species (e.g. see [3]). Compilations of repeat content from published sources reveal positive correlations between genome size and repeat proportions [4] and less consistent relationships with the proportion of TEs [5] over a range of eukaryote genome sizes. However, comparing repeat composition and dynamics between species is not straightforward because of different methods used to identify and characterize repeats (e.g. sequencing platforms, bioinformatic tools and similarity thresholds used, see Supplementary Fig. 1 and Supplementary Table 1 for comparison of estimates here with published estimates). Furthermore, most studies have focussed on species with small and medium sized genomes (< 10 Gbp/1C, where 1C-value is the amount of DNA in the gametic nucleus), limiting our understanding of repeat composition and dynamics across the full spectrum of genome sizes.

Here, using the same analytical approach, we analyse the repeat content of 101 species that encompass much of the enormous ~2,400 range in genome size diversity encountered in seed plants (0.063 - 88.55 Gbp/1C), which is comparable to the known range for diploid species (0.086-100.1 Gbp/1C, Supplementary Table 2, Extended Data Fig. 1). We focus on cytological diploids and exclude polyploids (except for two species, to include the smallest plant genomes), so that our results reflect the evolutionary history of repeat dynamics, rather than being complicated by the recent history of polyploidy (see Methods). Where possible we have used publicly available Illumina genomic sequence data. However, because these are dominated by plants with small and medium sized genomes, we also generated Illumina sequence data for a further 22 species (20 angiosperms and two gymnosperms) to extend the range of genome sizes included in our dataset. The complete dataset comprises 89 angiosperms (two early-diverging species, 63 eudicots and 24 monocots) and 12 gymnosperms (one cycad, two gnetophytes, *Ginkgo* and eight conifers) (Supplementary Table 3A).

Using the same parameters in all-to-all sequence comparisons, we grouped total genomic DNA sequences (Illumina reads) from each species into four categories based on the number of mutual similarity hits: (i) sequences present in ≤ 20 copies/1C genome, containing genes, associated non-coding regions and uncharacterized sequences, (ii) low copy repeats (21-500 copies/1C), (iii) medium copy repeats (501-10,000 copies/1C), and (iv) high copy repeats (> 10,000 copies/1C) (see Methods). We also analysed the abundance of conserved (retro)transposon protein coding domains using methods in Neumann et al. [6]. Data were analysed using linear modelling, applying linear, quadratic and cubic terms to explore significant ($p < 0.0001$) shifts in repeat dynamics with genome size (see

Supplementary Tables 4 and 5). See Methods for further information on sources of data, genome size estimates, detection of (retro)transposons and statistical approaches, .

When we plotted the repetitive proportion of the genome (genome proportion) of all repeats present in > 20 copies per genome against genome size (Fig. 1a, Extended Data Fig. 2), we observed that the slope of the graph varied across the range of genome sizes analysed. For species with small genomes (up to ~5 Gbp/1C) there was a steep and broadly linear increase in the proportion of the genome that is repetitive with genome size (from ~9% to ~70%). Beyond ~5 Gbp/1C, the slope of the graph asymptoted at a genome proportion of repeats of ~80%. However above ~10 Gbp/1C, the slope of the graph started to decline significantly ($p < 0.001$; Supplementary Table 4), so that the species with the largest genome size analysed (i.e. *Viscum album*, 88.55 Gbp/1C) had a genome proportion of repeats of only 55%. When the same data were plotted to show how repeats accumulate, a curvilinear relationship was revealed (Fig. 1b). The plot thus shows how the total amount of repeats (in Gbp) rises more slowly with genome size in species with large genomes than in species with small genomes.

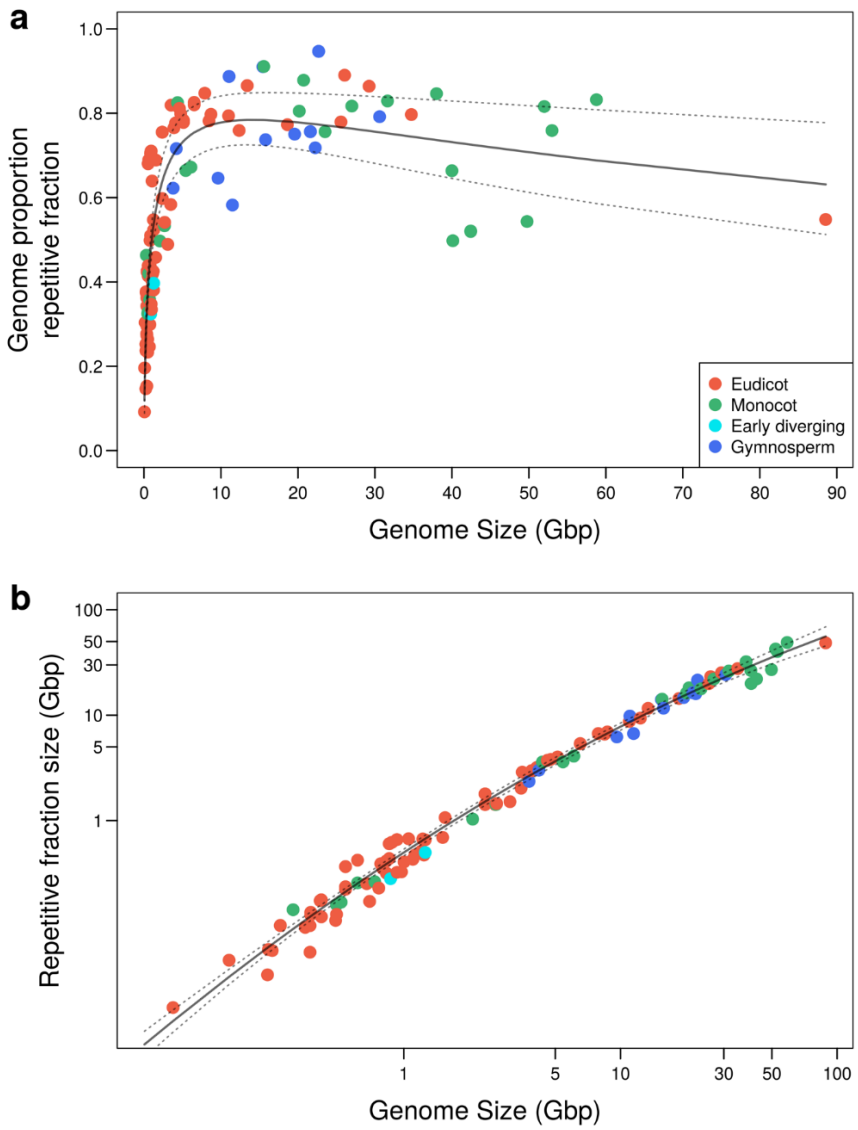


Fig. 1 Content of repeats present in > 20 copies in the genomes of 101 seed plant species ranging in size from 0.063 - 88.55 Gbp/1C and encompassing much of the known range of genome sizes encountered in seed plants. (a) Genome proportion plotted against genome size. Note how the graph profile does not asymptote near a repeat genome proportion of 1, as might be expected. Instead

above ~10 Gbp/1C the slope of the graph changes and the proportion of the genome that is repetitive in > 20 copies declines. The 99% confidence intervals are shown by dotted lines. (b) The size of the repetitive fraction in Gbp in the genome plotted against genome size. Note on this log scale how the fitted line follows a curvilinear relationship.

We also fitted the data to consider any phylogenetic non-independence in the datasets (using phylogenetic generalized least square models (PGLS), see Supplementary Tables 6-8). When the models are fitted with an Ornstein–Uhlenbeck process, the shapes of the curves remain similar (Extended Data Fig. 2). However, there is limited phylogenetic diversity in the upper region of genome sizes and the shape of the curves for the largest genome sizes is not recovered in all models (e.g. using Brownian motion). We particularly lack data for eudicots above 35 Gb/1C, however, surprisingly, given the diversity of eudicots, outside of *Viscum* there is only one species currently known with a larger genome than 35 Gbp/1C, and that species (*Hepatica nobilis*) is a tetraploid [7]. We note however, that the recently sequenced genome of the Mexican axolotl (*Ambystoma mexicanum* - 32 Gbp/1C) also has a lower proportion of repeats (~61%) than might be expected given its large genome [8], suggesting similar trends in animals (see also below).

The changing abundance of repeats across the range of genome sizes was also reflected in the proportion of genome represented by (retro)transposon protein coding domains, which varied from close to zero in species with some of the smallest genome sizes, up to ~12% in species with genome sizes ~5-10 Gbp/1C and then typically declining in species with genome sizes above ~10 Gbp/1C, a relationship also recovered in PGLS analyses (Supplementary Table 7, Extended Data Fig. 3).

For species with small and medium sized genomes (up to ~10 Gbp/1C) the linear increase in repeat genome proportion with genome size (Fig. 1a) was associated with a marked and significant ($p < 0.0001$) decrease in the genome proportion of sequences present in ≤ 20 copies and an increase in the proportion of higher copy repeats, particularly middle copy repeats ($p < 0.0001$, Extended Data Fig. 2, Supplementary Table 4). Repeats in species with genome sizes in this range are reported to be turning over rapidly, with half-lives of just tens of thousands [9] to a few million years (e.g. [10]). In addition, such genomes are characterised by having a relatively small number of specific repeats occupying a large proportion of the genome. For example, in *Vicia pannonica*, a specific family of Ty3/gypsy Ogre retrotransposons comprises ~38% of the genome [11]. Such data and comparisons between related species suggest that repeats in this genome size range are turning over rapidly, with sequence mutations and changes in repeat copy numbers (through amplification and deletion), leading to homogenous repeats that are divergent between species.

In stark contrast, for species with larger genomes (> ~10 Gbp/1C, Fig. 1a, Extended Data Fig. 2a), the genome proportion of single and low copy (≤ 20 copies) sequences significantly increases ($p < 0.0001$) with genome size. This is accompanied by a significant decrease ($p < 0.0001$) in the genome proportion of middle copy repeats (Extended Data Fig. 2d), although there remained a significant ($p < 0.0001$) increase in the genome proportion of higher copy repeats, with the slope of the regression being higher for eudicots compared with monocots (Extended Data Fig. 4d, h). It is likely that these observations arise through substantial increases in the quantities of degraded repeats, rather than an increase in the number of genes and/or gene regulatory regions. Degraded repeats arise from point mutations, indels and rearrangements, and they may be so substantial that they render repeats into tracks of unique or low copy sequences. If amplified repeats are not excised at the rate that they accumulate, they will become fossilised in the genome and mutate to non-repetitive DNA. We suggest that this dynamic substantially influences all species with large genomes and hence the differences in genome dynamics reported between angiosperms and gymnosperms [12,13] are not due to their contrasting phylogenetic histories, but in fact reflect their contrasting genome sizes (Fig. 2). This is further supported by studies showing that the large genomes of the lungfish *Neoceratodus forsteri* (52 Gbp/1C) [14] and five species of salamander (15-44 Gbp/1C) [8,15], are also comprised of a large

collection of heterogenous repeats, indicating that this contrasting repeat dynamic of large genomes also occurs in animals.

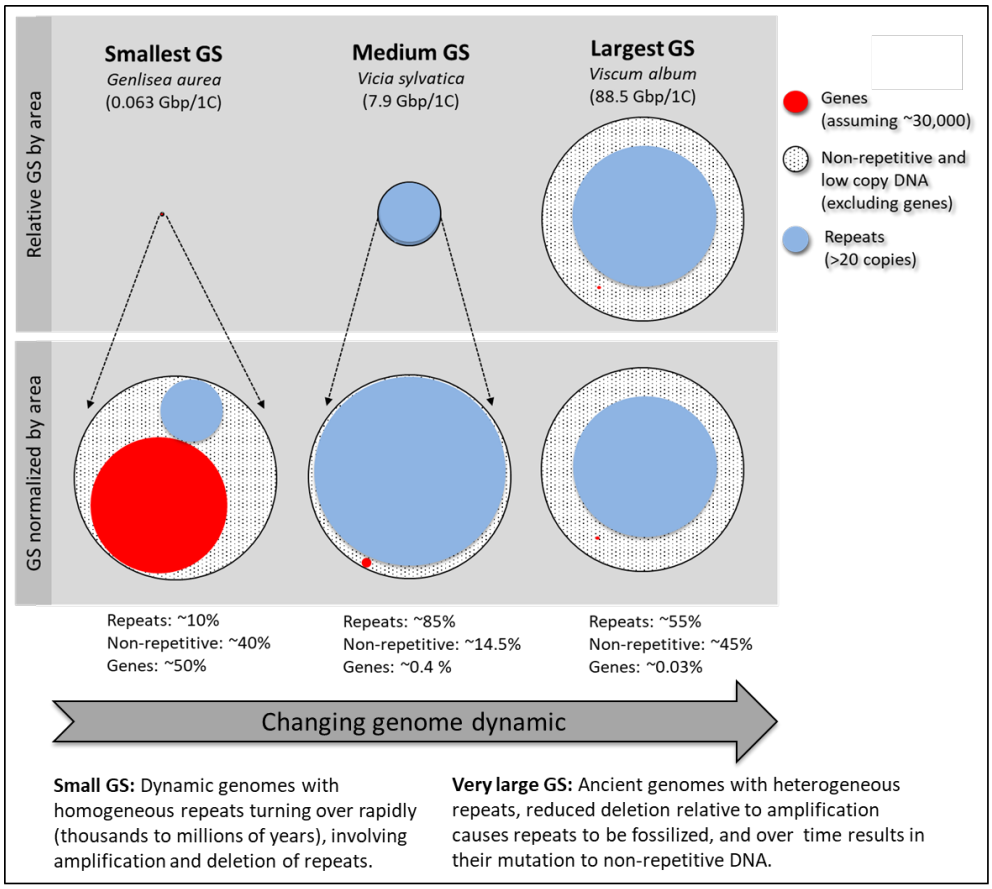


Fig. 2. Diagrammatic illustration of contrasting repeat dynamics across the range of seed plant genome size (GS), scaled by area, and by proportions of genes (protein coding genes, excluding introns) assuming ~30,000 genes (excluding genes from (retro)transposons), each of 1 Kbp (excluding introns), non-repetitive and low copy DNA excluding genes (i.e. uncharacterized sequences, gene regulatory region and introns), and repeats (> 20 copies).

DNA repair, especially non-homologous end joining, has been implicated in playing a role in the rate at which DNA is removed from the genome, with size of deletions being larger in the small genome of *Arabidopsis thaliana* compared with the larger genome of barley [16]. Recombination-based processes are known to remove repeats from the genome, with genome size being influenced by the rate of sequence elimination. For example, analyses of linkage map data indicate that plants with larger genomes have reduced recombination rates compared with those possessing smaller genomes [17], while the rate of recombination is generally negatively correlated with (retro)transposon abundance [18]. Epigenetic silencing of (retro)transposable elements may play a role by reducing the frequency of recombination-based removal of repeats and hence creating a positive feedback loop between repeat accumulation and recombination suppression [18]. Certainly, it can be envisaged that in species with large genomes, the slow degradation of repeats, putatively trapped within regions of low recombination, could lead to ever increasing accumulation of low and single copy sequences [19,20] and hence give rise to the repeat profiles observed here in species with large genomes.

Given this fundamental shift in genome dynamics observed here it is predicted that diploid species with genomes larger than ~10 Gbp/1C are indeed on a “one-way ticket to genomic obesity” [21] with limited options for a “return ticket”. This shift in the evolutionary trajectory of the genome will in turn impact a diversity of physiological processes which are associated with larger genomes, such as the

higher metabolic and nutrient costs needed to build and maintain large genomes [22] and the longer cell cycle times [23]. Furthermore, because DNA occupies a volume, larger genomes will have larger nuclei and cells, impacting nuclear [24] and cellular physiology (e.g. water retention, gas exchange in photosynthesis, ad flowering time, [25,26,27]). Overall, these impacts will ultimately play a role in constraining how and where plants grow [28], their ability to tolerate extreme environmental conditions [29,30] and hence their long term survival over evolutionary time.

Methods

Selection of plant species for analysis. We analysed genomic DNA from 129 seed plants comprising 101 species (see Supplementary Table 3A). All except two species are chromosomally diploid. However, we also included *Genlisea aurea* and *G. nigrocaulis* which have been shown to occur in a tetraploid clade of *Genlisea* that includes species showing a dysploid chromosome number series, and which might indicate that their genomes are partially diploidised [31]. These two *Genlisea* species are included because their genomes are amongst the smallest so far reported for seed plants. Nevertheless, all angiosperm and many gymnosperm lineages are considered to have undergone polyploidy or whole genome duplication (WGD) events in their evolutionary history, often on multiple occasions, since the predicted WGD at the base of seed plants [32,33].

Whilst we had ample data to resolve trends occurring at the lower end of the genome size range, at the upper end of the genome size range we were restricted in the data available by our requirement to analyse diploid species, so that our work was not confounded by recent polyploidy events. There are, however, few diploid species with genomes sizes >20 Gbp/1C and fewer still with available sequence data for us to analyse. Nevertheless, we selected representatives from all the key lineages known to us that have large genome sizes (>20 Gbp/1C), to maximise our ability to search for phylogenetic independence in the trends across the known range of genome sizes. The sources of the material used for analysis, the person who supplied the materials and the dates collections were made are provided in Supplementary Table 3 B.

Estimation of genome repetitiveness. To estimate the genome proportion and copy number of repetitive components of the nuclear genomes, we first filtered out all low-quality reads, reads containing adapter sequences and reads with similarity to the plastid and mitochondrial genomes. Reads were then trimmed to uniform length, with all reads in a sample trimmed to between 90-100 nt in length. The pre-processing was performed using *single_fastq_filtering.R* to remove sequences which did not pass the quality threshold (quality at least 10 in 95% of all bases). The program for filtering is available in the git repository (https://bitbucket.org/repeatexplorer/re_utilities). A random sample of filtered and trimmed forward reads was used in all analyses to give 5% coverage (0.05x) of the nuclear genome, meaning that we analysed read numbers so that the sequencing depth was the same for all species, irrespective of genome size, which varied from 0.063 - 88.55 Gbp/1C. Use of only forward reads eliminates similarity hits between paired end reads, which would distort the results.
For each species we compared reads using an all-to-all similarity search with the optimized BLASTn program mgblast as implemented in the TGI Clustering Tool (https://sourceforge.net/projects/tgicl/ version 2.1-1). The following mgblast command line options were used: -W18 -UT -X40 -KT -JF -F “mD” -v100000000 -b100000000 -D4 -C50 -H30. Based on our long experience with RepeatExplorer [34], we tested two similarity thresholds, one with 90% identity (RepeatExplorer default) and another with 80% identity. The analysis on a subset of the analysed species showed the approach used to detect repeats in > 20 copies is robust. See Supplementary Fig.2 comparing thresholds at 80% and 90%.

For the final analysis, we chose to use 80% identity threshold which provides better sensitivity towards divergent repeats. The method will detect repeats of any size, with equal probability, based only on

the similarity between individual sequence reads (80% over at least 55% of the read length). The approach finds repeats, irrespective of the repeat length, which can range in length from tens of bases to kilobases.

For each read we counted the number of similarity hits. Based on the number of similarity hits, reads were divided into bins (groups). The number of reads which did not produce any similarity hits corresponds to the fraction of the genome with ≤ 20 copies of a sequence. This group contains genes, associated non-coding regions and uncharacterized sequences. The number of reads which produced 1-25, 26-500 and more than 500 similarity hits corresponds to the size of the fraction of the genome with copy numbers of 21-500, 501-10,000 and > 10,000 copies, respectively. In other words, the proportion of reads in each group is equivalent to the proportion of the different repetitive fractions in the genome.

Flow cytometry and genome size data. Genome size data were taken from the Plant DNA C-values database release 7.1 (https://cvalues.science.kew.org/)[35], Ickert-Bond et al. [36] or measured here; in each case the source reference is provided in Supplementary Table 3 B. For three species (*Trillium ovatum*, *Hyacinthoides non-scripta*, *Fritillaria cirrhosa*) new values were determined from fresh leaf material using flow cytometry and propidium iodide (PI) staining of the nuclei, following best practice methods described in Pellicer and Leitch [37]. Briefly, about 1 cm^2 of freshly harvested leaf material of the study species and the calibration standard species were co-chopped in 2ml of nuclei isolation buffer (Pellicer and Leitch [37]) and filtered through 30 μm nylon mesh filter and stained by adding in 100 μl of 1 mg/ml PI solution. Nuclei were then incubating for about 30 min on ice. The genome sizes were estimated using a CyflowSL Partec flow cytometer fitted with a 100 mW green (532 nm) solid state Cobalt Samba laser and the resulting flow histograms were analysed using Partec software (FloMax version 2.7). On analysis, only peaks with a CV <2.5 were considered suitable for estimating genome sizes. Polygon gating was applied only to exclude debris outside G1 fluorescence range after visual inspection of nuclei populations. See Supplementary Fig. 3 to show an example of the gating approach. A minimum of 5,000 nuclei were measured for each analysis. Genome sizes were estimated by analysing the PI mean peak position of the calibration standard species of known genome size and the mean peak position of the nuclei of the experimental material. To obtain genome size estimates, a minimum of three samples and three replicate runs were analysed for each species.

Detection of (retro)transposon coding sequences. The identification of (retro)transposable element protein domains was performed using the REXdb reference database [6], which includes conserved polyprotein domain sequences extracted from 80 species representing the major groups of green plants (Viridiplantae). To estimate the contribution of (retro)transposons to genome size, we used REXdb and similarity searches performed using DIAMOND version 0.9.13 (https://github.com/bbuchfink/diamond) with the following parameter settings: --max-target-seqs 1 --min-score 30 --freq-sd 1000.

Statistical analyses. Two datasets were analysed:
(A) Copy number data, with the genome divided into four categories comprising: (i) sequences with ≤ 20 copies per 1C genome (containing genes, associated non-coding regions and uncharacterized sequences); (ii) low copy repeats (sequences with 21-500 copies); (iii) middle copy repeats (sequences with 501-10,000 copies), and; (iv) high copy repeats (sequences with >10,000 copies). This dataset comprised all 129 individuals from 101 species. Mean values were estimated for species represented by more than one individual, for a total of 89 angiosperms (2 early-diverging angiosperms, 63 eudicots and 24 monocots) and 12 gymnosperms. (B) (Retro)transposon data for 77 species comprising 69 angiosperms (1 early-diverging angiosperm, 53 eudicots, 15 monocots), and eight gymnosperms.

The genome proportion occupied by the different categories of repeats (see above) were analysed with a beta regression [38], using R version 3.3.3 with the R package *betareg* [39] version 3.1-0. This method

is similar to a logistic regression, but rather than being restricted to a binary variable, it allows a continuous variable bounded within a (0, 1) interval to be fitted as a dependent variable. It does not allow values of 0 and 1. High copy repeats were absent in nine species, therefore this variable was subsequently transformed by: $(y \cdot (n - 1) + 0.5)/n$, where n is the sample size [40]).

Beta regression consists of two sub-models, where the first part, the location model, predicts the mean and is estimated by a logit link. The second part is the precision model with a log link, which returns a *phi* coefficient. The higher the *phi*, the higher the precision (and the lower the dispersal, or variance). We first assessed whether to include a polynomial term in genome size predicting the genome proportion of a repetitive element. Models were fitted with and without a non-orthogonal polynomial term; the contribution of the polynomial term was assessed by a combination of diagnostic plots, a log-likelihood test to assess model specification, and comparison of the AICs. We analysed the associations between genome proportions and repetitive elements for all species in the datasets. We analysed the associations between genome proportions and repetitive elements for all species in the datasets. We also performed a second analysis with plant clade (eudicot, monocot, and gymnosperm) as a factor variable to test for differences between these clades. Early-diverging angiosperms were removed at this point of analysis because of the very small sample size (i.e. n = 1, n = 2), a sample size too small to meaningfully analyse as a clade. We tested whether the slope of the regression line was significantly different between these clades by including an interaction between genome size and plant clade. Fitted regressions were assessed with a combination of diagnostic plots of residuals and outliers, including standardized weighted residual ("sweighted2") plots recommended by Cribari-Neto and Zeileis [39], the likelihood-ratio test of squared linear predictors to test for model misspecification, the Breusch-Pagan test against heteroscedasticity, and AIC for comparing models. We also tested incorporating further regressors (plant higher group or genome size) in the *phi* sub-model as precision parameters but found these were not necessary to account for e.g. heteroscedasticity.

In all regressions, genome size was natural-log (ln) transformed to account for the left skewed distribution and wide variation (from 0.063 to 88.55 Gbp/1C) in this variable. We applied this process to both the copy number and the (retro)transposon datasets. Variances between clades were sufficiently similar to enable between clade analyses using the stated statistical tests.

We also analysed the associations between the genome proportion of repeats and genome size within a phylogenetic context. We pruned the Daphne phylogenetic tree [41] to include the species in our dataset (Supplementary Fig. 4). Tips for taxa within clades that are absent from this phylogeny (e.g. *Gnetum gnemon*, *Trillium*) were manually added, and polytomies were transformed to dichotomies with the *ape* package [42] version 5.0. Proportional branch lengths were applied to the phylogeny with FigTree [43] version 1.4.3. Phylogenetic signal was estimated using Blomberg's K and Pagel's lambda with the *phytools* package [44] version 0.6-44. To account for phylogenetic non-independence and for the curvilinear trends in the data, we fitted phylogenetic generalized least square models (PGLS) with an Ornstein–Uhlenbeck process and with Brownian motion using the *gls* function from the *nlme* package [45] version 3.1-131 . As with the beta regressions, PGLS models were fitted with non-orthogonal polynomial terms, and we assessed whether second/third order polynomial terms were appropriate. We used a phylogeny with proportional branch lengths, and for comparison of the effects of the phylogeny, we also fitted a PGLS with a phylogeny transformed to a cladogram.

Reporting summary

Further information on research design is available in the Nature Reporting Summary linked to this paper.

Data availability

Data are available in two databases: (1) The genomic DNA data analysed were available in the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home) or Illumina sequenced here before archiving the sequences on ENA (Supplementary Table 3). Details of the ENA accession identifier for each sample are available in Supplementary Table 3 A. Details of the source of the plant material and sequencing platform are given in Supplementary Table 3 A. (2) Genome size data were taken from reported estimates given in the Plant DNA C-values database release 7.1 (https://cvalues.science.kew.org/) or from source publications not yet included in the database; in each case the source reference is provided in Supplementary Table 3 A, see column 'S' in Supplementary Table 3 A) and the species analysed here are listed in Table 3 B (see the 'Methods' section 'Flow cytometry and genome size data' for further information).

Code availability

Most of the code used to analyse these data are integral to the published, established program packages as stated above, with the parameter settings given, as appropriate. For filtering out all low-quality sequence reads, reads containing adapter sequences and reads with similarity to the plastid and mitochondrial genomes new code was generated and this is available in the git repository https://bitbucket.org/repeatexplorer/re_utilities.

References

1 Lisch, D. How important are transposons for plant evolution? *Nature Rev. Genet.* **14**, 49-61 (2013).

2 Bennetzen, J. L. & Park, M. Distinguishing friends, foes, and freeloaders in giant genomes. *Curr. Opin. Genet. Dev.* **49**, 49-55 (2018).

3 Kersey, P. J. Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.* **48**, 1-8 (2019).

4 Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Phil. Trans. Roy. Soc. B: Biol. Sci.* **370** (2015).

5 Elliott, T. A. & Gregory, T. R. Do larger genomes contain more diverse transposable elements? *BMC Evol. Biol.* **15**, 69 (2015).

6 Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* **10**, 1 (2019).

7 Mabuchi, T., Kokubun, H., Mii, M. & Ando, T. Nuclear DNA content in the genus *Hepatica* (Ranunculaceae). *J. Plant Res.* **118**, 37-41 (2005).

8 Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50-55 (2018).

9 Stritt, C., Wyler, M., Gimmi, E. L., Pippel, M. & Roulin, A. C. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass *Brachypodium distachyon*. 10.1111/nph.16308 (2019).

10 Ma, J. X. & Bennetzen, J. L. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc. Nat. Acad. Sci. USA* **103**, 383-388 (2006).
11 Neumann, P., Koblížková, A., Navrátilová, A. & Macas, J. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. *Genetics* **173**, 1047-1056 (2006).
12 Nystedt, B. et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579-584 (2013).
13 De La Torre, A. R., Li, Z., Van de Peer, Y. & Ingvarsson, P. K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363-1377 (2017).
14 Metcalfe, C. J., Filée, J., Germon, I., Joss, J. & Casane, D. Evolution of the Australian lungfish (*Neoceratodus forsteri*) genome: A major role for CR1 and L2 LINE elements. *Mol. Biol. Evol.* **29**, 3529-3539 (2012).
15 Sun, C., López Arriaza, J. R. & Mueller, R. L. Slow DNA loss in the gigantic genomes of salamanders. *Genome Biol. Evol.* **4**, 1340-1348 (2012).
16 Vu, G. T. H., Cao, H. X., Reiss, B. & Schubert, I. Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol.* **214**, 1712-1721 (2017).
17 Tiley, G. P. & Burleigh, J. G. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evol. Biol.* **15**, 194 (2015).
18 Kent, T. V., Uzunović, J. & Wright, S. I. Coevolution between transposable elements and recombination. *Phil. Trans. Roy. Soc. B: Biol. Sci.* **372**, 20160458 (2017).
19 Maumus, F. & Quesneville, H. Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS ONE* **9**, e94101 (2014).
20 Kelly, L. J. et al. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol.* **208**, 596-607 (2015).
21 Bennetzen, J. L. & Kellogg, E. A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**, 1509-1514 (1997).
22 Leitch, A. R. & Leitch, I. J. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629-646 (2012).
23 Francis, D., Davies, M. S. & Barlow, P. B. A strong nucleotypic effect of DNA C-value on the cell cycle regardless of ploidy level. *Ann. Bot.* **101**, 747-757 (2008).
24 Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: The Impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* **180**, 1-52 (2019).
25 Roddy, A. B. et al. The scaling of genome size and cell size limits maximum rates of photosynthesis with implications for ecological strategies. *Int. J. Plant Sci.* **181**, 75-87 (2020).
26 Lawson, T. & Blatt, M. R. Stomatal size, speed, and responsiveness impact on photosynthesis and water use efficiency. *Plant Physiol.* **164**, 1556-1570 (2014).
27 Franks, P. J. & Beerling, D. J. Maximum leaf conductance driven by CO_2 effects on stomatal size and density over geologic time. *Proc. Natl. Acad. Sci. USA* **106**, 10343-10347 (2009).
28 Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. J. Genome size diversity and its impact on the evolution of land plants. *Genes* **9**, 88 (2018).

29 Knight, C. A., Molinari, N. A. & Petrov, D. A. The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* **95**, 177-190 (2005).
30 Vidic, T., Greilhuber, J., Vilhar, B. & Dermastia, M. Selective significance of genome size in a plant community with heavy metal pollution. *Ecol. Appl.* **19**, 1515-1521 (2009).
31 Fleischmann, A. et al. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* **114**, 1651-1663 (2014).
32 Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Rev. Genet.* **18**, 411-424 (2017).
33 Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348-363 (2018).
34 Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792-793 (2013).
35 Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301-305 (2019).
36 Ickert-Bond, S. M. et al. Polyploidy in gymnosperms – Insights into the genomic and evolutionary consequences of polyploidy in *Ephedra*. *Mol. Phyl. Evol.* **147**, 106786 (2020).
37 Pellicer, J. & Leitch, I. J. in *Molecular Plant Taxonomy* Vol. 1115 *Methods in Molecular Biology (Methods and Protocols)* (ed P. Besse) Ch. 14, 279-307 (Humana Press, Totowa, NJ, 2014).
38 Ferrari, S. & Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Statistics* **31**, 799-815 (2004).
39 Cribari-Neto, F. & Zeileis, A. Beta Regression in R. *J. Statistical Software* **34**, 1-24 (2010).
40 Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Meth.* **11**, 54-71 (2006).
41 Durka, W. & Michalski, S. G. Daphne: a dated phylogeny of a large European flora for phylogenetically informed ecological analyses. *Ecology* **93**, 2297-2297 (2012).
42 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).
43 Rambaut, A. FigTree v. 1.4.3 [Internet]. http://tree.bio.ed.ac.uk/software/figtree. (2012).
44 Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217-223 (2012).
45 Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1*, http://cran.r-project.org/package=nlme (2017).

Acknowledgements

We thank NERC (NE/G020256/1), the Czech Academy of Sciences (RVO:60077344) and Ramón y Cajal Fellowship (RYC-2017-2274) funded by the Ministerio de Ciencia y Tecnología (Gobierno de España) for support. In addition, the work was supported by the ERDF/ESF project ELIXIR-CZ - Capacity building

(No. CZ.02.1.01/0.0/0.0/16_013/0001777) and ELIXIR CZ research infrastructure project (LM2015047) for the access to computing and storage facilities. We also thank NERC for funding a studentship to Steven Dodsworth and the China Scholarship Council for funding Wencai Wang. Finally, we thank Jeannine Marquardt for supplying DNA of *Hyacinthoides non-scripta*.

Author Contributions

ARL, IJL, J Macas and P Novák conceived the experiment, and designed, implemented and coordinated the project. P Novák conducted genomic sequence analysis, PN conducted (retro)transposon protein coding domains analysis, JP and JM provided material and flow cytometry analysis and MSG statistical analysis. JM, LJK, SD, WW, A Kovařík, AK and JP provided sequence data and experimental advice. All authors were involved in writing the manuscript.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Information

Supplementary Tables

Supplementary Table 1. The data and methods used in previously published work to estimate repeat genome proportions (GPs) are provided in an Excel spreadsheet (filename: Novak_Supplementary_Table_1 (2 Sept).xlsx).

Supplementary Table 2. The range of genome sizes (GS/1C) and genome proportions (GP) occupied by all repeat sequences present in > 20 copies for the different groups of plants analysed here. The full list of species analysed together with their GS and GP are available in Supplementary Table 3.

Plant Group	Number of species analysed	Minimum GS	Maximum GS	Minimum repetitive GP (GS)	Maximum repetitive GP (GS)
Angiosperms					
Early-diverging angiosperms	2	0.87 Gbp (*Amborella trichopoda*)	1.26 Gbp (*Persea borbonia*)	32% In: *Amborella trichopoda* (0.87 Gbp)	40% In: *Persea borbonia* (1.26 Gbp)
Eudicots	63	0.06 Gbp (*Genlisea aurea*)	88.55 Gbp (*Viscum album*)	9% In: *Genlisea aurea* (0.06 Gbp)	89% In: *Podophyllum peltatum* (26.06 Gbp)
Monocots	24	0.31 Gbp (*Zostera marina*)	58.78 Gbp (*Paris quadrifolia*)	32% In: *Oryza sativa* (0.49 Gbp)	92% In: *Agapanthus africanus* (15.57 Gbp)
Gymnosperms	12	3.79 Gbp (*Gnetum gnemon*)	30.61 Gbp (*Encephalartos ferox*)	58% In: *Gingko biloba* (11.49 Gbp)	95% In: *Pinus sylvestris* (22.69 Gbp)

Supplementary Table 3. Details of the materials used in sequencing and repeat genome proportions (GP) and genome size (GS) data: (A) shows the 101 plant species analysed for total repeat GP and the GPs of each of the four repeat categories based on the number of mutual similarity hits. It also shows the GPs of transposable elements (TEs), genome sizes (GS, bp/1C) of the species analysed and the sources of that data; (B) shows the species in which the GS data were obtained in this work, and; (C) lists the technical and biological replicates examined with the sources of the data (filename: Novak_Supplementary_Table_3 (2 Sept).xlsx).

Supplementary Table 4. Beta regression output testing association between *ln*-transformed genome size (GS) and genome proportions of (a) all repeats (i.e. > 20 copies per 1C genome), (b) sequences in ≤ 20 copies (including genes, associated non-coding regions and uncharacterised sequences), (c) low copy repeats (21-500 copies), (d) middle copy repeats (501-10,000 copies), and (e) high copy repeats (≥ 10,000 copies). Shown for each category of repeat is the output of the beta regression with all species (n=101) shown in Supplementary Fig. 3, and of the beta regression in which the clades (eudicot, monocot, and gymnosperm) are included (n=99). The *phi* coefficients refer to the precision model with a log link; a higher *phi* indicates a lower dispersion (variance). The baseline level (intercept) is the eudicot clade. LR is the log-likelihood, DF = degrees of freedom, and the pseudo R^2 is a measure of the overall variation explained by the model. GS^2 and GS^3 are the quadratic and cubic terms included in the regression to assess any curvilinear trends (see also Supplementary Information 6). Below each sub-table is the Breusch-Pagan (BP) test against heteroscedasticity, where the null hypothesis of homoscedasticity is rejected if the p-value is < 0.05. The p-values associated with regression coefficients are two-tailed.

a) All repeats (copy number > 20), pseudo R^2= 0.7162, LR = 83.64 on 5 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-0.067	-0.209	0.074	0.072	-0.933	0.3509
GS	0.730	0.612	0.847	0.060	12.165	< 0.0001
GS^2	0.002	-0.060	0.064	0.032	0.064	0.9490
GS^3	-0.031	-0.049	-0.014	0.009	-3.520	0.0004
phi	16.577	12.111	21.043	2.279	7.275	< 0.0001

BP = 4.9823, df = 3, p-value = 0.1731

With higher group: pseudo R^2= 0.7193, LR= 82.83 on 7 DF

	Estimate	CI 2.5%	CI 97.5%	Std. error	z-value	p-value
Intercept	-0.026	-0.175	0.123	0.076	-0.341	0.7330
GS	0.742	0.616	0.868	0.064	11.549	< 0.0001
GS^2	0.001	-0.063	0.064	0.032	0.018	0.9857
GS^3	-0.031	-0.049	-0.013	0.009	-3.309	0.0009
Gymnosperm	-0.127	-0.504	0.250	0.192	-0.661	0.5088
Monocot	-0.120	-0.394	0.154	0.140	-0.857	0.3916
phi	16.792	12.221	21.364	2.332	7.200	< 0.0001

BP = 6.5117, df = 5, p-value = 0.2596

b) Sequences ≤ 20 copies, pseudo R^2= 0.7162, LR = 83.64 on 5 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	0.067	-0.074	0.209	0.072	0.933	0.3509
GS	-0.730	-0.847	-0.612	0.060	-12.165	< 0.0001

GS^2	-0.002	-0.064	0.060	0.032	-0.064	0.9490
GS^3	0.031	0.014	0.049	0.009	3.520	0.0004
phi	16.577	12.111	21.043	2.279	7.275	< 0.0001

BP = 5.0895, df = 3, p-value = 0.1654

With higher group: pseudo R^2= 0.7193, LR= 82.83 on 7 DF

	Estimate	CI 2.5%	CI 97.5%	Std. error	z-value	p-value
Intercept	0.026	-0.123	0.175	0.076	0.341	0.7330
GS	-0.742	-0.868	-0.616	0.064	-11.549	< 0.0001
GS^2	-0.001	-0.064	0.063	0.032	-0.018	0.9857
GS^3	0.031	0.013	0.049	0.009	3.309	0.0009
Gymnosperm	0.127	-0.250	0.504	0.192	0.661	0.5088
Monocot	0.120	-0.154	0.394	0.140	0.857	0.3917
phi	16.792	12.221	21.364	2.332	7.200	< 0.0001

BP = 6.5118, df = 5, p-value = 0.2596

c) Low copy repeats pseudo R^2 = 0.07601, LR = 89.19 on 4 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-1.054	-1.194	-0.915	0.071	-14.815	< 0.0001
GS	0.134	0.025	0.243	0.056	2.399	0.0164
GS^2	-0.039	-0.077	-0.001	0.019	-2.029	0.0424
phi	16.703	12.190	21.216	2.303	7.254	< 0.0001

BP = 3.1318, df = 2, p-value = 0.2089

With higher group: pseudo R^2 = 0.1172, LR = 89.34 on 6 DF

	Estimate	CI 2.5%	CI 97.5%	Std. error	z-value	p-value
Intercept	-1.089	-1.237	-0.942	0.075	-14.512	< 0.0001
GS	0.104	-0.008	0.216	0.057	1.821	0.0686
GS^2	-0.037	-0.075	0.001	0.019	-1.898	0.0577
Gymnosperm	0.408	0.068	0.748	0.174	2.354	0.0186
Monocot	0.034	-0.248	0.315	0.144	0.236	0.8138
phi	17.345	12.608	22.083	2.417	7.176	< 0.0001

BP = 12.941, df = 4, p-value = 0.01157

d) Middle copy repeats, pseudo R^2 = 0.5687, LR = 100.4 on 4 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-1.607	-1.779	-1.436	0.088	-18.333	< 0.0001
GS	0.633	0.483	0.782	0.076	8.297	< 0.0001
GS^2	-0.109	-0.155	-0.063	0.024	-4.612	< 0.0001
phi	14.593	10.582	18.603	2.046	7.132	< 0.0001

BP = 3.7616, df = 2, p-value = 0.1525

With higher group: pseudo R^2 = 0.5709, LR = 98.42 on 6 DF

	Estimate	CI 2.5%	CI 97.5%	Std. error	z-value	p-value
Intercept	-1.576	-1.755	-1.397	0.091	-17.261	< 0.0001
GS	0.666	0.512	0.821	0.079	8.436	< 0.0001
GS^2	-0.114	-0.162	-0.067	0.024	-4.706	< 0.0001
Gymnosperm	-0.263	-0.638	0.113	0.191	-1.371	0.1700
Monocot	-0.051	-0.361	0.260	0.159	-0.319	0.7500
phi	14.731	10.640	18.822	2.087	7.058	< 0.0001

BP = 5.3959, df = 4, p-value = 0.249

e) High copy repeats: pseudo R^2 = 0.5186, LR=158.7 on 3 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-2.737	-2.968	-2.506	0.118	-23.246	< 0.0001
GS	0.397	0.311	0.483	0.044	9.084	< 0.0001
phi	14.948	10.591	19.306	2.223	6.723	< 0.0001

BP = 7.4952, df = 1, p-value = 0.006186

With higher group: pseudo R^2 = 0.5067, LR = 157.6 on 5 DF

	Estimate	CI 2.5%	CI 97.5%	Std. error	z-value	p-value
Intercept	-2.687	-2.917	-2.457	0.118	-22.862	< 0.0001
GS	0.492	0.396	0.589	0.049	9.996	< 0.0001
Gymnosperm	-0.371	-0.797	0.055	0.217	-1.708	0.0876
Monocot	-0.592	-0.953	-0.231	0.184	-3.218	0.0013
phi	16.896	11.932	21.860	2.533	6.671	< 0.0001

BP = 6.144, df = 3, p-value = 0.1048

Supplementary Table 5. Beta regression output testing association between genome proportions of total (retro)transposable elements (TE) and *ln*-transformed genome size (GS). (**a**) all species (n=77) and in higher clades (eudicots, monocots and gymnosperms) (n = 76). The baseline level (intercept) is the eudicot clade. In (**b**), the same analyses are shown but with the exclusion of *Sorghum bicolor* which was shown to be an extreme outlier in a QQ plot of weighted residuals. LR is the log-likelihood, DF = degrees of freedom, and the pseudo R^2 is a measure of the overall variation explained by the model. GS^2 is the quadratic term included in the regression to assess a curvilinear trend. Below each sub-table is the Breusch-Pagan (BP) test against heteroscedasticity, where the null hypothesis of homoscedasticity is rejected if the p-value is < 0.05. See also Supplementary Information 6, and Supplementary Fig. 4.. The p-values associated with regression coefficients are two-tailed.

a) Transposable elements (n=77): pseudo R^2 = 0.2289, LR = 174 on 4 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-2.602	-2.722	-2.483	0.061	-42.735	< 0.00001
GS	0.350	0.237	0.462	0.057	6.093	< 0.00001
GS^2	-0.090	-0.127	-0.054	0.019	-4.84	< 0.00001
phi	93.830	63.990	123.673	15.230	6.163	< 0.00001

BP test = 0.372, df = 2, p-value = 0.8304

With higher clade: TEs (n=76), pseudo R^2 = 0.4437, LR = 188.4 on 6 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-2.564	-2.678	-2.450	0.058	-44.057	< 0.00001
GS	0.349	0.242	0.455	0.054	6.407	< 0.00001
GS^2	-0.081	-0.116	-0.045	0.018	-4.481	< 0.00001
Gymnosperm	-0.054	-0.337	0.230	0.145	-0.37	0.7113
Monocot	-0.348	-0.597	-0.099	0.127	-2.738	0.00619
phi	105.620	71.842	139.405	17.240	6.128	< 0.00001

BP test= 4.3266, df = 4, p-value = 0.3636

b) *Sorghum bicolor* outlier removed:

Transposable elements (n=76): pseudo R^2 = 0.4228, LR = 189.4 on 4 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-2.553	-2.645	-2.461	0.047	-54.363	< 0.00001
GS	0.322	0.235	0.409	0.044	7.259	< 0.00001

GS^2	-0.090	-0.118	-0.061	0.015	-6.163	< 0.00001
phi	158.100	107.663	208.544	25.740	6.143	< 0.00001

BP test = 0.029472, df = 2, p-value = 0.9854

With higher clade: TEs (n=75), pseudo R^2 = 0.4437, LR = 188.4 on 6 DF

	Estimate	CI 2.5%	CI 97.5%	Std. Error	z value	p-value
Intercept	-2.539	-2.631	-2.448	0.047	-54.484	< 0.00001
GS	0.320	0.234	0.406	0.044	7.270	< 0.00001
GS^2	-0.090	-0.119	-0.061	0.015	-6.115	< 0.00001
Gymnosperm	0.065	-0.169	0.298	0.119	0.544	0.58700
Monocot	-0.062	-0.263	0.139	0.102	-0.605	0.54500
phi	165.310	112.232	218.393	27.080	6.104	< 0.00001

BP test = 6.7367, df = 4, p-value = 0.1505

Supplementary Table 6. (a) Phylogenetic signal in *ln*-genome size and copy number within 101 species, estimated with Blomberg's K-statistic, Pagel's lambda, and their corresponding p-values. Phylogenetic signal was tested in a phylogeny with proportional branch lengths (P), and without branch lengths (cladogram C). (b) shows phylogenetic signal in total (retro)transposable elements in 77 species. Lambda is estimated with a maximum likelihood approach; λ = 0 represents a star phylogeny which indicates independent evolution of the trait; λ = 1 indicates a strong correlation between species as expected under Brownian evolution. A significant lambda p-value, obtained from the likelihood ratio test, indicates a strong phylogenetic signal. K is a scaled ratio of the observed trait variance among species over the contrasts variance expected under Brownian motion. A significant (one-tailed) p-value for K indicates that closely related species are more similar to each other than random pairs of species.

(a)

Trait	Phylo	K	K p-value	Lambda (λ)	lambda p-value
ln-genome size (n=101)	P	0.112	0.001	0.901	< 0.00001
	C	0.463	0.001	0.875	< 0.00001
Sequences ≤ 20 copies	P	0.079	0.001	0.863	< 0.00001
	C	0.332	0.001	0.818	< 0.00001
Low copy repeats (21-500 copies)	P	0.028	0.666	0.147	0.0195
	C	0.128	0.646	0.152	0.0212
Middle copy repeats (501-10,000 copies)	P	0.061	0.001	0.784	< 0.00001
	C	0.262	0.001	0.721	< 0.00001
High copy repeats (> 10,000 copies)	P	0.060	0.001	0.844	< 0.00001
	C	0.254	0.001	0.763	< 0.00001

(b)

Trait	Phylo	K	K p-value	lambda	lambda p-value
ln-genome size (n=78)	P	0.210	0.001	0.942	< 0.00001
	C	0.710	0.001	0.912	< 0.00001
Total transposable elements	P	0.072	0.004	0.751	0.0002
	C	0.273	0.002	0.676	0.00002

Supplementary Table 7. PGLS summary, fitted with an Ornstein–Uhlenbeck process: (a-e) copy number regressed on *ln*-genome size (GS); and (f) total (retro)transposable elements regressed on *ln*-GS. A phylogenetic tree with proportional branch lengths was used to infer evolutionary relationships. A curvilinear association was estimated with a quadratic term (GS^2). In contrast to the beta regression, a cubic term was not significant in sequences with ≤ 20 copies (a, b). Sample size n=101 species in copy number repeats, and n=77 species in total (retro)transposable elements. We also used a phylogeny with branch lengths transformed to a cladogram, but results were so similar that they are not shown in this table (see Supplementary Figures 3-4). The p-values obtained from the PGLS are two-tailed.

a) All repeats (copy number > 20)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.484	0.449	0.519	0.018	27.222	< 0.0001
GS	0.165	0.139	0.192	0.014	12.040	< 0.0001
GS^2	0.000	-0.014	0.014	0.007	-0.022	0.9822
GS^3	-0.007	-0.011	-0.003	0.002	-3.452	0.0008

b) Sequences ≤ 20 copies

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.516	0.481	0.551	0.018	29.041	< 0.0001
GS	-0.165	-0.192	-0.139	0.014	-12.040	< 0.0001
GS^2	0.000	-0.014	0.014	0.007	0.022	0.9822
GS^3	0.007	0.003	0.011	0.002	3.452	0.0008

c) Low copy repeats (21-500 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.260	0.233	0.287	0.014	19.003	< 0.0001
GS	0.025	0.005	0.045	0.010	2.471	0.0152
GS^2	-0.008	-0.015	-0.001	0.004	-2.130	0.0357

d) Middle copy repeats (501-10,000 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.188	0.156	0.220	0.016	11.572	< 0.0001
GS	0.073	0.049	0.096	0.012	6.091	0.0000
GS^2	-0.010	-0.019	-0.002	0.004	-2.482	0.0148

e) High copy repeats (> 10,000 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value

Intercept	0.059	0.037	0.802	0.011	5.382	< 0.0001
GS	0.038	0.028	0.484	0.005	7.544	< 0.0001

f) Total (retro)transposable elements

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.068	-2.722	-2.483	0.004	17.363	< 0.0001
GS	0.020	0.237	0.462	0.003	7.037	< 0.0001
GS^2	-0.001	-0.127	-0.054	0.002	-0.455	0.6504
GS^3	-0.001	63.990	123.673	0.000	-2.779	0.0069

Supplementary Table 8. PGLS summary, fitted with Brownian motion: (a-e) copy number regressed on *ln*-genome size (GS); and (f) total (retro)transposable elements regressed on *ln*-GS. A phylogenetic tree with proportional branch lengths was used to infer evolutionary relationships. A curvilinear association was estimated with a quadratic term (GS^2). In contrast to the beta regression, a cubic term was not significant in sequences with ≤ 20 copies (a, b). Sample size n=101 species in copy number repeats, and n=77 species in total (retro)transposable elements. We also used a phylogeny with branch lengths transformed to a cladogram, but results were so similar that they are not shown in this table. The p-values from the PGLS are two-tailed.

a) All repeats (copy number > 20)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.467	-0.129	1.062	0.304	1.535	0.1279
GS	0.131	0.099	0.163	0.016	8.001	< 0.0001
GS^2	-0.016	-0.029	-0.004	0.006	-2.687	0.0085

b) Sequences ≤ 20 copies

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.533	-0.062	1.129	0.304	1.755	0.0823
GS	-0.131	-0.163	-0.099	0.016	-8.001	< 0.0001
GS^2	0.016	0.004	0.029	0.006	2.687	0.0085

c) Low copy repeats (21-500 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.252	-0.365	0.869	0.315	0.801	0.4253
GS	0.060	0.027	0.093	0.017	3.556	0.0006
GS^2	-0.016	-0.029	-0.004	0.006	-2.534	0.0129

d) Middle copy repeats (501-10,000 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.184	-0.388	0.756	0.292	0.632	0.5288
GS	0.058	0.027	0.089	0.016	3.690	0.0004
GS^2	-0.010	-0.022	0.001	0.006	-1.732	0.0864

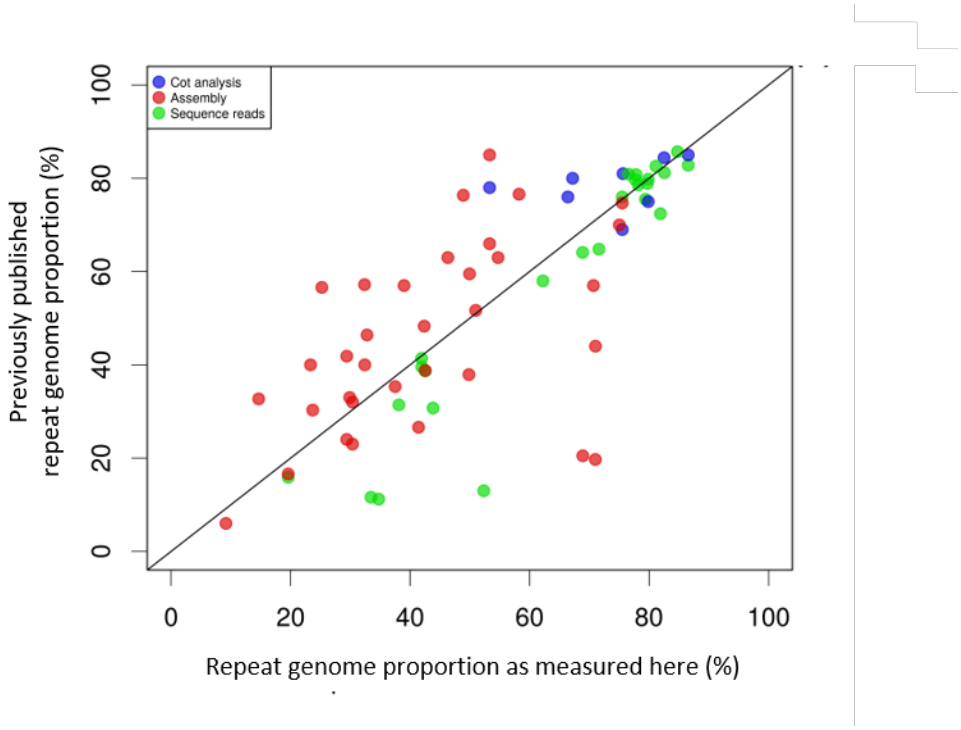
e) High copy repeats (> 10,000 copies)

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.030	-0.379	0.439	0.209	0.144	0.8858
GS	0.013	-0.009	0.035	0.011	1.125	0.2631
GS^2	0.010	0.002	0.018	0.004	2.331	0.0218

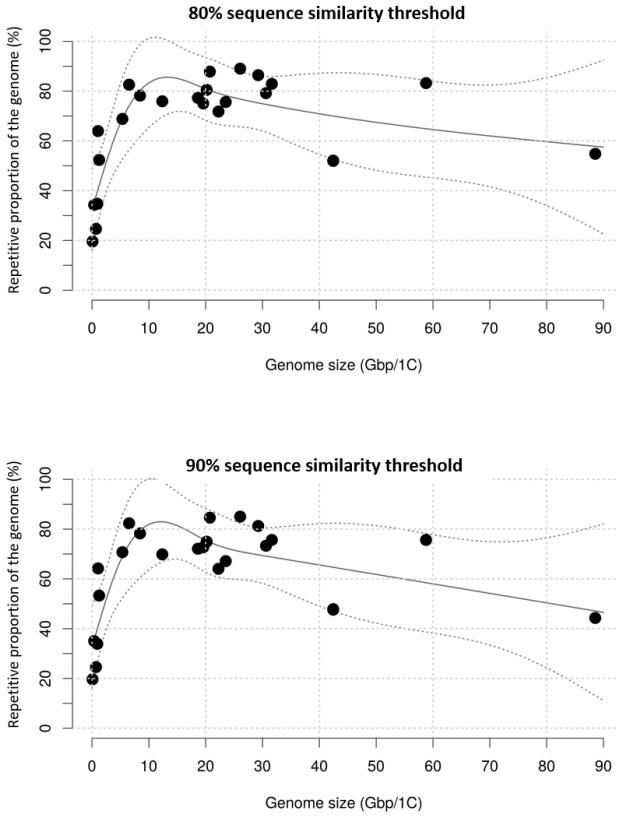
f) **Total (retro)transposable elements**

	Estimate	CI 2.5%	CI 97.5%	Std. Error	t-value	p-value
Intercept	0.067	-0.031	0.164	0.050	1.336	0.1855
GS	0.014	0.008	0.021	0.003	4.425	< 0.0001
GS^2	-0.004	-0.006	-0.001	0.001	-3.222	0.0019

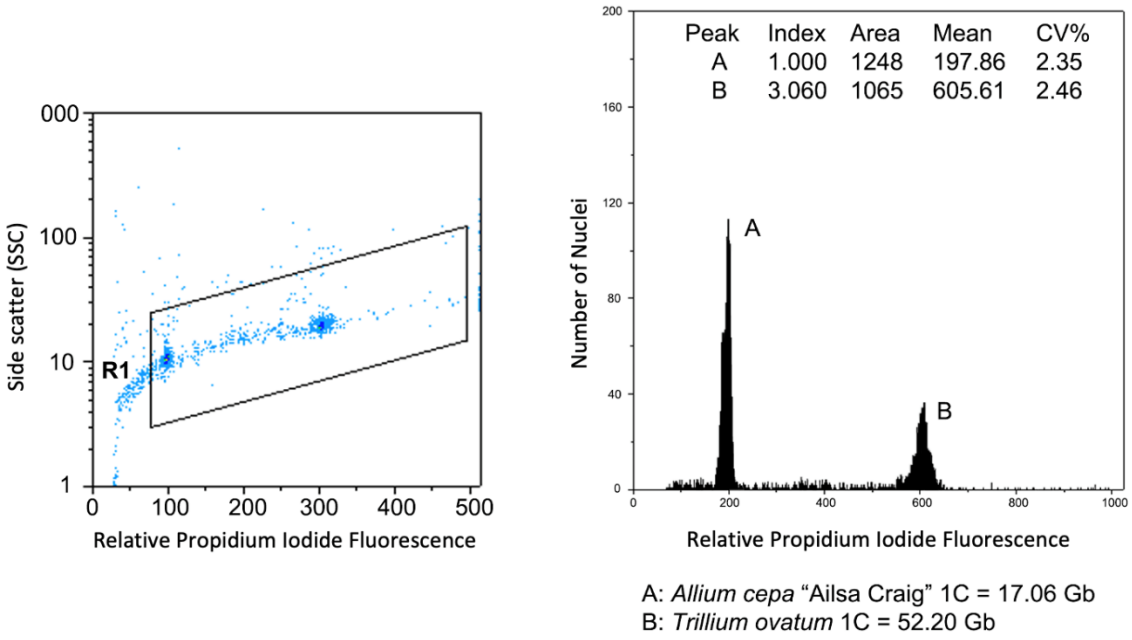
Supplementary Figures



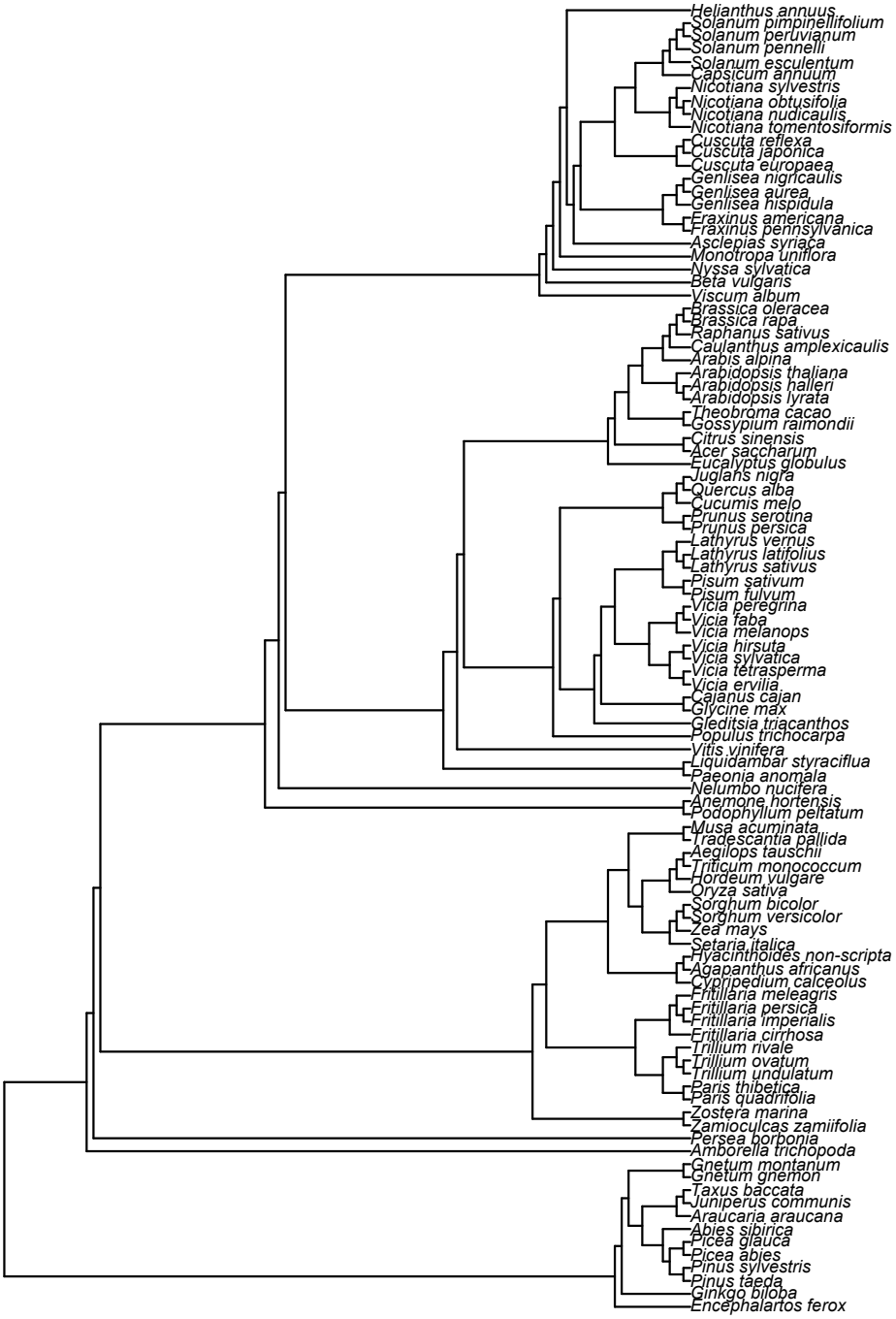
Supplementary Fig. 1. Comparison of estimates of repeat genome proportion (GP) as measured here with previously published estimates. The figure shows the proportion of repeats with > 20 copies estimated here versus repeat genome proportions obtained from previously published estimates for 54 species analysed. The previously published estimates are divided into three categories depending on the analytical method: (i) Cot analysis include estimates based on reassociation kinetics of DNA; (ii) assembly, based on the estimates from annotation of whole genome assemblies; (iii) sequence reads based on low pass genomic sequencing. Whilst there is a clear overall relationship between data sets, as expected, there are also considerable discrepancies for some species between our estimates and those published previously, with previous estimates being both larger and smaller than those reported here, revealing the importance of a uniform approach to estimate repeat genome proportion. Values used for plotting and source references are given in Supplementary Table 1.



Supplementary Fig. 2. A comparison on the repetitive proportion of the genome at two different sequence similarity thresholds (80 and 90 percent identity (PID)) to determine their effects in predicting repetitive sequence genome proportion across the range of plant GS analysed.



Supplementary Fig. 3. Flow cytometry analysis of nuclei of *Trillium ovatum* with the calibration standard *Allium cepa*. The plot on the left shows PI fluorescence intensity and side scatter (SSC) values for each nucleus. The nuclei analysed are in the gated area (R1). The plot on the right shows a flow histogram of the nuclei in the gated area. Peak positions enable calculation of the genome size of *Trillium ovatum* to be estimated.



Supplementary Fig. 4. Phylogenetic tree of the 101 species in the repeats copy number dataset.