The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies

Introduction

In recent decades, interest in plant genome size (i.e. the total amount of DNA in the unreplicated haploid nucleus; Greilhuber et al., 2005) has been growing exponentially as the biological, evolutionary and ecological significance of this key biodiversity trait is increasingly recognized (e.g. see reviews by Greilhuber & Leitch, 2013; Pellicer et al., 2018). Such interest is no doubt, in part, underpinned by the staggering diversity of genome sizes encountered within land plants (e.g. especially angiosperms which show a range of c. 2400-fold; Pellicer et al., 2010) and the considerable diversity in some algal clades, with the most variable being in the Chlorophyta clade of green algae, which have a range of 274-fold. Certainly, it is now clear that genome size can have an impact at many scales, from influencing gene and genome dynamics (e.g. Dodsworth et al., 2015) to playing a role at the whole-plant level, influencing, for example, plant growth strategies, plant community composition, plant–animal interactions, evolutionary trajectories and ecosystem dynamics (e.g. Suda et al., 2015; Guignard et al., 2016, 2019; Simonin & Roddy, 2018). In the era of fast-evolving high-throughput-sequencing technologies, C-values also provide baseline information necessary for estimating whole-genome sequencing costs, as these are directly linked to the size of the genome (e.g. Li & Harkess, 2018).

Accessing plant genome size data online

Although genome size data have been accumulating in the literature since 1951 when the first estimate for a plant was published (Ogur et al., 1951), for many years such data remained scattered across the literature or compiled into lists available only in hard copy (e.g. Bennett & Smith, 1976; Bennett et al., 2000). This made the search to determine whether a particular genome size was available slow and tedious. Release 1 of the Plant DNA C-values database in 2001 contributed significantly to overcoming such a bottleneck and helped to revolutionize the field by facilitating large-scale comparative phylogenetic analyses across different plant groups (e.g. Leitch & Bennett, 2002; Soltis et al., 2003). Since then, six updates have been released, culminating in the latest, which went live in April 2019 (Leitch et al., 2019) and which collates data from 1067 original publications and personal communications.

The Plant DNA C-values database (release 7.1, April 2019)

The new release of the Plant DNA C-values database (https://cvalues.science.kew.org/) contains genome size data for 12 273 species – increasing the number of species represented by 44% (i.e. an additional 3763 new species) since the last update in 2012 (Bennett & Leitch, 2012; Garcia et al., 2014). The vast majority of data are for angiosperms, with estimates for 10 770 species. However, the database also contains C-values for all other major land plant groups, with data for 421 gymnosperms, 303 pteridophytes (comprising 246 ferns (monilophytes) and 57 lycophytes) and 334 bryophytes (209 mosses, 102 liverworts and 23 hornworts). Data are also available for 445 ‘algae’, comprising species belonging to evolutionarily distinct higher-order lineages (i.e. Rhodophyta, Chlorophyta, and the streptophyte green algae within Kingdom Plantae, and Phaeophyta and Heterokonta within the Stramenopiles).

The new user-friendly interface of the database has a range of searching and output options so that the user can extract and display specific information as required. For example, queries can be made using the whole database, or limited to specific taxonomic lineages and levels (e.g. families, genera). In addition, more detailed searches can be made by specifying certain criteria (e.g. restricting searches to defined genome size ranges, ploidal levels, chromosome numbers, higher taxonomic groups, life cycles etc.). The new release also includes predictive typing in the taxonomy boxes (i.e. family, genus and species levels). Although the species names given in the original publications are always kept, the family affiliations for the angiosperms follow those of the most recent Angiosperm Phylogeny Group (APG IV) update (Angiosperm Phylogeny Group, 2016).

Where more than one estimate has been reported for a species, by default the output of a query will just display the prime genome size estimate, which represents the most consistent value obtained under best-practice methods (as originally defined by Bennett & Smith, 1976). Nevertheless, all data published for a given species in the database can be accessed by changing the search parameters to display ‘all estimates’.

Data curation and best practices for estimating nuclear DNA contents

Although we encourage efforts to generate novel data in underexplored groups, we are now committed to include, wherever possible, only C-values that have been estimated using best-practice protocols (e.g. Doležel et al., 2007; Pellicer & Leitch, 2014). In brief, the original publications (or authors) should provide information about the method used for genome size estimation (with flow cytometry currently recommended as best practice), the
number of replicates made, a clear indication of the quality of measurements (e.g. coefficient of variation (CV%) of the peaks in the flow cytometry histograms), the intercalating fluorochrome used (e.g. propidium iodide), and the calibration standard selected and its genome size used to convert relative measurements into absolute DNA C-values.

As genome size estimations made using flow cytometry do not provide cytological information (e.g. chromosome number, ploidal level, presence of B chromosomes etc.), the database does not infer this information if it is not reported in the original reference source. Although this means that many genome size estimates are not accompanied by cytological data (i.e. there are no cytological data for 3701 species), this deliberately avoids the risk of misinterpreting the genome size data, particularly for species with multiple cytotypes (Kolár et al., 2017). Nevertheless, users of the Plant DNA C-values database can opt to check publicly available databases such as the Chromosome Counts Database (CCDB) (Rice et al., 2015) to evaluate if chromosome counts have been made for a particular species, and whether multiple cytotypes have been reported. Certainly, users are warned that without prior cytological knowledge of a given taxonomic group, interpretation of the chromosome number or ploidal level based solely on a genome size estimation can be very misleading, as highlighted by Suda et al. (2006).

Growth, gaps and novel insights in plant genome size diversity data

Release 7.1 of the Plant DNA C-values database has seen a significant increase of data, averaging a rate of c. 537 estimates per year (between 2012 and 2019) for species not previously included in the database. When it comes to taxonomic groups and phylogenetic representation, angiosperms still make up the bulk of new data (as in previous releases), comprising 85.7% of the 3228 entries newly added to the database (Fig. 1). This extends our knowledge of C-values at the species level up to c. 3% of the currently recognized diversity (based on the most recent revised estimate of 369,434 extant species of angiosperms; see Nic Lughadha et al., 2016). At the higher taxonomic levels, representation currently stands at 15% of genera (2118 out of c. 14,000), as well as 63% of families (262 out of 416) and 94% of orders (58 out of 62) recognized by APG IV (2016), illustrating the need to continue to prioritize efforts to fill taxonomic gaps, particularly at the genus and family levels, in future research.

There are two major lineages beyond angiosperms whose representation in the database has improved significantly from the previous release, namely monilophytes (increasing the number of species with data from 101 to 246), and algae (increase from 253 to 445 species). Nevertheless, given the number of species recognized in these two groups (> 11,000 monilophytes and > 20,000 algae), phylogenetic representation is still extremely low, especially for many of the monophyletic groups of algae.

With respect to monilophytes, not only has the taxonomic coverage increased, but the new data now include the first reported example of extreme genomic obesity beyond angiosperms. This arises from the Hidalgo et al. (2017b) report of a giant genome in the whisk-fern Tmesipteris obliqua, whose genome size of 146,500 Mb/1C is over twice the size of the previous record holder (i.e. 71,221 Mb in Psilotum nudum), thereby extending the range of C-values in monilophytes from 94-fold to 196-fold). This new
value closely rivals the gigantic genome of the angiosperm Paris japonica (1C = 148 851 Mb; Pellicer et al., 2010), which is the largest so far reported for any eukaryote (Hidalgo et al., 2017a; Fig. 2). Another relatively unexplored group from a genome size perspective is the algae, which comprise several evolutionarily distinct lineages, some with closer affinities to animals than plants (e.g. those belonging to Heterokonta, such as diatoms, and the brown algae (Phaeophyta)). They are often neglected because of the methodological challenges they pose when estimating genome size (Voglmayr, 2007; Mazalov et al., 2011). Evolutionary relationships among many algal lineages are still controversial, but recent phylogenetic studies provide strong support for three of the streptophyte algal lineages (i.e. Charophyceae, Coleochaetophyceae and Zygnematophyceae) forming a monophyletic group that is sister to land plants (Embryophyta) (Wickett et al., 2018). Thus, data on the size of these algal genomes are critical for providing insights into the evolutionary implications of genome size diversity before the colonization of land. Despite data still being sparse (i.e. six Charophyceae, four Coleochaetophyceae and 49 Zygnematophyceae species), the new estimates in the database highlight considerable genome size variation. While Charophyceae (1882–19 208 Mb/1C) and Coleochaetophyceae (343–1348 Mb/1C) have genomes that fall within the range encountered in bryophytes (i.e. 156–19 560 Mb/1C), which include the sister group to all vascular plants, larger genomes are found in Zygnematophyceae, including the largest for any algal lineage reported so far: the polyploid desmid Microaerystias rotata (= 31 723 Mb/1C). Such a large genome extends beyond those so far encountered in bryophytes and highlights that the potential for genome size expansion is not restricted to vascular land plants (Fig. 2). At the other end of the scale, the Chlorophyta and Rhodophyta include the smallest genomes reported for any free-living photosynthetic organism (12.46 Mb/1C in the green chlorophyte Ostreococcus tauri and 13.20 Mb/1C in the rhodophyte Galliera sulphuraria; these genome sizes fall within the values reported for some bacterial genomes, and are considered to represent the bare limits of life for free-living photosynthetic eukaryotes; Peers & Niyogi, 2008) (Fig. 2).

### Prospects and future directions for the Plant DNA C-values database to achieve comprehensive understanding of plant genome size diversity

As the relevance of genome size continues to be recognized across a diversity of research fields, there are growing opportunities for conducting increasingly large-scale comparative analyses to enhance understanding of the evolutionary and ecological significance of the immense genome size diversity in plants. Yet, to ensure such analyses are robust, there is clearly an ongoing need to generate novel, high-quality genome size data and to make them accessible, especially targeting some of the more poorly represented lineages in the Plant DNA C-values database to make the data more comprehensive of plant diversity as a whole.

It seems likely that flow cytometry will continue to be the main technique for generating reliable genome size data in plants, given that there is now a good understanding of the diversity of factors that influence this and robust recommendations for best-practice approaches (Doležel et al., 2007). Nevertheless, with the rapid advances in DNA-sequencing technology, accompanied by the increasing rate at which whole-genome sequences are published (Kersey, 2019), several bioinformatic tools are now being developed that aim to estimate genome size directly from whole-genome data. Some of these are based on the analysis of k-mer frequency distributions in the sequence dataset (e.g. GenomeScope or findGSE; Sun et al., 2017; Vurture et al., 2017). More recently, a mapping-based genome size estimation approach has been reported (MGSE; Pucker, 2019) that infers genome size from short-read sequencing data by mapping reads to a highly contiguous assembly, and assuming a random fragmentation of DNA when preparing it for sequencing (i.e. equal distribution of reads over the complete sequence).

Although these bioinformatic methods might appear to be promising alternatives to flow cytometry, it is noted that they often give genome size estimates that are lower than those from flow cytometry. The causes of such bias are currently unclear. Further, it is not known whether such approaches are even applicable to the many plants which are polyploid, highly repetitive and/or are particularly large, as all these features are likely to present challenges to the underpinning assumptions of the bioinformatic pipelines (Sun et al., 2017; Vurture et al., 2017). Instead, these observations highlight the urgent need for appropriate comparative analyses to be conducted across the diversity of plant genomes which compare bioinformatic and flow cytometry estimates determined from the same specimen. Only then will we be able to shed light on the

---

![Fig. 2 Range and mean (±) genome sizes (1C-values) encountered in the different lineages represented in the Plant DNA C-values database (release 7.1, April 2019). The number in brackets following the name of each lineage corresponds to the number of species with genome size data.](image-url)
underlying causes generating the discrepancies, and hence their reliability. Given these potential limitations, genome size estimates derived from such methods have not yet been included in the Plant DNA C-values database. Nevertheless, we will continue to monitor the development of these and related approaches for consideration in future releases of the Plant DNA C-values database.

Acknowledgements

We thank Emmeline Johnston for her help in collating and entering all the genome size data into the database, as well as all those authors who have kindly provided data through their publications that have been incorporated into the database. We are also grateful to Eduardo Toledo and his IT team at the Royal Botanic Gardens, Kew for their help in upgrading the webpage interface and enabling the new release to go live.

Author contributions

JP and IJL contributed equally to conceiving and writing this letter.

ORCID

Ilia J. Leitch https://orcid.org/0000-0002-3837-8186
Jaume Pellicer https://orcid.org/0000-0001-7632-9775

Jaume Pellicer and Ilia J. Leitch*

Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, TW9 3DS, UK (*Author for correspondence: tel +44 208 332 5329; email i.leitch@kew.org)

References


**Key words:** algae, C-values, flow cytometry, land plants, plant DNA C-values database.

Received, 16 August 2019; accepted, 8 October 2019.