frontiers | Frontiers in Plant Science

TYPE Methods
PUBLISHED 14 March 2024
DOI 10.3389/fpls.2024.1328966

OPEN ACCESS

EDITED BY
George V. Popescu,
Mississippi State University, United States

REVIEWED BY
Ruslan Kalendar,
University of Helsinki, Finland
Boas Pucker,
Technical University of Braunschweig,
Germany

*CORRESPONDENCE
France Denoeud
fdenoeud@genoscope.cns.fr

RECEIVED 27 October 2023
ACCEPTED 22 February 2024
PUBLISHED 14 March 2024

CITATION
Guenzi-Tiberi P, Istace B, Alsos IG, The PhyloNorway Consortium, Coissac E, Lavergne S, The PhyloAlps Consortium, Aury J-M and Denoeud F (2024) LocoGSE, a sequence-based genome size estimator for plants. *Front. Plant Sci.* 15:1328966. doi: 10.3389/fpls.2024.1328966

LocoGSE, a sequence-based genome size estimator for plants

Pierre Guenzi-Tiberi[1], Benjamin Istace[1], Inger Greve Alsos[2], The PhyloNorway Consortium, Eric Coissac[3], Sébastien Lavergne[3], The PhyloAlps Consortium, Jean-Marc Aury[1] and France Denoeud[1*]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France, [2]The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, Norway, [3]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA (Laboratoire d'Ecologie Alpine), Grenoble, France

Extensive research has focused on exploring the range of genome sizes in eukaryotes, with a particular emphasis on land plants, where significant variability has been observed. Accurate estimation of genome size is essential for various research purposes, but existing sequence-based methods have limitations, particularly for low-coverage datasets. In this study, we introduce LocoGSE, a novel genome size estimator designed specifically for low-coverage datasets generated by genome skimming approaches. LocoGSE relies on mapping the reads on single copy consensus proteins without the need for a reference genome assembly. We calibrated LocoGSE using 430 low-coverage Angiosperm genome skimming datasets and compared its performance against other estimators. Our results demonstrate that LocoGSE accurately predicts monoploid genome size even at very low depth of coverage (<1X) and on highly heterozygous samples. Additionally, LocoGSE provides stable estimates across individuals with varying ploidy levels. LocoGSE fills a gap in sequence-based plant genome size estimation by offering a user-friendly and reliable tool that does not rely on high coverage or reference assemblies. We anticipate that LocoGSE will facilitate plant genome size analysis and contribute to evolutionary and ecological studies in the field. Furthermore, at the cost of an initial calibration, LocoGSE can be used in other lineages.

KEYWORDS

genome size estimation, genome size, ploidy, genome-skimming, environmental DNA, plant genomics, 1C, 1Cx

1 Introduction

Genome size is a trait that has been shown to vary greatly between eukaryotes, but does not correlate with organismal complexity (Mirsky and Ris, 1951; Cavalier-Smith, 1978; Gregory, 2005). Notably, in land plants, there is a, 2400-fold variation of genome size between different species (Pellicer et al., 2018), which has been shown to be caused by lineage-specific insertion/excision dynamics of DNA elements such as retrotransposons (Bennetzen et al., 2005; Grover et al., 2008; Pellicer et al., 2018; Chase et al., 2023). For example, genome size increases have been related to retrotransposon invasions in Poaceae (Sanmiguel and Bennetzen, 1998; Hawkins et al., 2006; Piegu et al., 2006; Dai et al., 2022), Melanthiaceae (Pellicer and Leitch, 2014; Pellicer et al., 2021) or Gymnosperms (Morse et al., 2009; Ohri, 2021). In addition to retrotransposon invasions, giant genomes are thought to have arisen because of the lack of DNA removal (Kelly et al., 2015). Besides, whole genome duplications are frequent in plants (Jaillon et al., 2007; Jiao et al., 2012, Jiao et al., 2014; Murat et al., 2017; Ren et al., 2018) and contribute to a lesser extent to genome size variations (Pellicer et al., 2018). They also result in variable ploidy levels between plant species as well as inside populations (Weiss-Schneeweiss and Schneeweiss, 2013).

Interest in plant genome size is high not only because of the need to estimate sequencing efforts required to obtain full genome sequences (Kelly et al., 2012; Li and Harkess, 2018; Pellicer and Leitch, 2020) but more importantly because this trait has been shown to be of evolutionary and ecological significance (Greilhuber and Leitch, 2013; Pellicer et al., 2018; Blommaert, 2020). The Kew Plant DNA C-values Database (Pellicer and Leitch, 2020); https://cvalues.science.kew.org/) is a valuable resource for plant genome sizes and provides C-values (i.e. total amount of DNA in the unreplicated haploid nucleus, or holoploid genome size (Greilhuber et al., 2005) for more than 12,000 plant species. These measures are generally obtained by flow cytometry (Dolezel et al., 2007; Sliwinska et al., 2022; Temsch et al., 2022), an experimental technique that usually requires live or frozen tissues with intact cells. Such requirements are not always easy to fulfill, for instance for botanists who work with (sometimes ancient) herbarium collections. Sequencing data provide an interesting alternative to estimate genome size (Pflug et al., 2020). Indeed, genome skimming approaches aimed at obtaining plant phylogenetic barcodes have been expanding over the last decade (Coissac et al., 2016; Li et al., 2019; Nevill et al., 2020; Fu et al., 2022). These approaches rely on low coverage short-read sequencing (usually less than 10 Million Illumina read pairs) in order to assemble chloroplastic genomes (or targeted barcode genes) and were shown to be applicable even on ancient herbarium samples (Alsos et al., 2020). Here, we present LocoGSE, a software to estimate monoploid genome size “1Cx” from such very low coverage datasets.

The monoploid genome size (1Cx) is the DNA content of the whole chromosome complement, irrespectively of the degree of polyploidy (Greilhuber et al., 2005). For diploid species, 1Cx is equal to 1C. Usually, genome size estimators do not specify which type of “genome size” they are estimating. In reality, all sequence-based genome size estimators are estimating monoploid (1Cx) rather than holoploid (1C) genome size. Such a distinction might not be essential for lineages that are mostly diploid [for instance insects (Pflug et al., 2020)], but when one wants to analyze plant genomes, where polyploidization events are frequent, it is important to have a clear definition of the genome size that is being estimated.

Previously described sequence-based methods for genome size estimation belong to two main categories: k-mer-based or mapping-based approaches (Sun et al., 2018; Liu et al., 2020; Pflug et al., 2020). K-mer-based approaches only require raw sequences but at a relatively high depth of coverage, usually above 30X for the most commonly used software GenomeScope (Vurture et al., 2017; Ranallo-Benavidez et al., 2020). These methods can not distinguish between two (or more) subgenomes with low degrees of heterozygosity, and will thus always predict the 1Cx genome size rather than 1C. In fact, GenomeScope 2.0 uses a ploidy level as input (default=2) and predicts a “genome haploid length” that actually corresponds to 1Cx and should be multiplied by the ploidy to obtain the 2C value (DNA content of a diploid cell). Hozza et al. designed a k-mer based approach (CovEst) for lower depths. Their tests showed promising results at depths of coverage as low as 1X but they were performed only on simulated genomes and a small bacterial (*E. coli*) genome (Hozza et al., 2015). The performance and optimal depth threshold for CovEst still need to be estimated on large and complex genomes. Pflug et al. showed its efficiency on various insects and three model organism species (*Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*), but the depths of coverage in their datasets were all over 30X (Pflug et al., 2020). Recently another package, RESPECT, was developed specifically for low-coverage genome skims. It was shown to perform better than CovEst to predict genome size from low coverage datasets, even at very low depth of coverage (0.5X) (Sarmashghi et al., 2021). However, as specified by the authors, RESPECT is designed and optimized to work with low coverage data, and should not be used with sequencing depths above 5X. That causes a problem when one wants to estimate the size of a genome without prior knowledge, since the genome size needs to be known in order to calculate the number of reads to use as input to the program. Finally, one needs to keep in mind that k-mer based methods are very sensitive to heterozygosity (Pucker, 2019) and thus need to be used cautiously. Current mapping-based approaches need to map the reads onto an assembly [ModEst (Pfenninger et al., 2022)] and some also necessitate a reference single copy gene set, like MGSE (Pucker, 2019) and Gnodes (Gilbert, 2022) for short reads, and Depthsizer (Chen et al., 2022) for long reads. These approaches imply that assembly and sometimes also annotation have been performed on the genome studied, which requires high sequencing depth. Consequently, they are not suited for very low coverage datasets, such as the ones produced by genome skimming projects.

Our genome size estimator, called LocoGSE (Low coverage based Genome Size Estimator), is a mapping-based approach that does not rely on a genome assembly, since the reads are mapped on a reference dataset of single copy genes (protein consensus) instead. Thus, it is particularly suitable for very low coverage short reads datasets. We calibrated LocoGSE using 430 Angiosperm low-coverage genome skimming datasets. Then, we performed a benchmark to compare its results and performances with other available genome size estimators, on plant datasets with various sequencing depths and properties (heterozygosity, ploidy). We show that monoploid genome size estimations made by LocoGSE are accurate even at very low coverage (<1X) and on highly heterozygous samples. Interestingly, the genome size predictions remain accurate at higher coverage, which allows its use without any prior knowledge about the size of the genome analyzed. Monoploid genome size estimations are also stable across individuals with varying ploidy levels.

2 Methods

2.1 Rationale

All sequence-based genome size predictors rely on the Lander-Waterman equation (Lander and Waterman, 1988), C = L/G, where G corresponds to genome size, L corresponds to cumulative length of sequenced nucleotides, C corresponds to sequencing depth of coverage. The aim is thus to estimate C in order to calculate G. The assumption behind LocoGSE and other mapping-based methods targeted on single-copy sequences (Pucker, 2019; Chen et al., 2022) is that the average depth on a set of single-copy sequences (usually single copy genes) is representative of the depth of coverage on the entire genome. Estimating the depth on a set of single-copy sequences should then allow us to estimate C and then G. However, it is important to take into account possible whole genome duplications (WGD) that may lead to various degrees of ploidy. If a tetraploidy event occurred recently, all the genome is duplicated, leading to a double genome size, but a similar proportion of single copy genes relative to the rest of the genome. At a given sequencing depth, the depth on single copy genes will then be the same for a recent tetraploid as for a diploid, and reflect the size of the monoploid genome, 1Cx (Greilhuber et al., 2005) (Figure 1). Consequently, genome size estimators based on mapping on single copy genes will always predict monoploid genome size (1Cx) rather than holoploid genome size 1C, as do k-mer based estimators. Conversely, the experimental protocol aiming at measuring the DNA content in cells, flow cytometry, is obviously measuring 1C genome sizes (Pellicer and Leitch, 2014). Thus, 1Cx genome size estimators have a promising application: they can be used to complement flow cytometry analysis, by estimating the ploidy level of plant specimens (by dividing 1C measurements by 1Cx estimations).

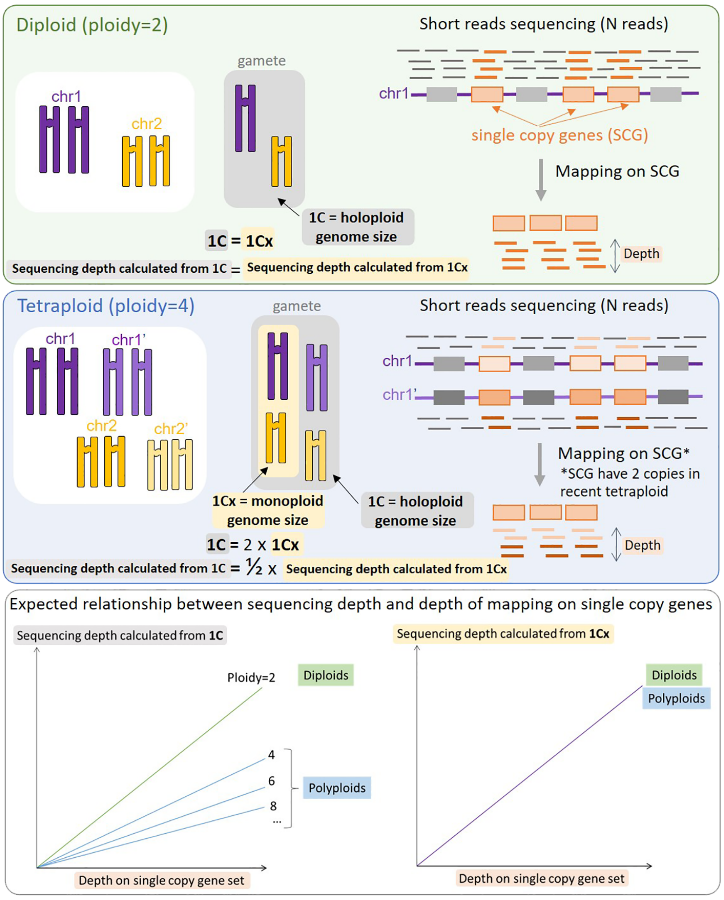
Diploid (ploidy=2)
Short reads sequencing (N reads)
gamete
chr1
chr2
chr1
single copy genes (SCG)
Mapping on SCG
1C = holoploid genome size
1C = 1Cx
Depth
Sequencing depth calculated from 1C = Sequencing depth calculated from 1Cx
Tetraploid (ploidy=4)
gamete
Short reads sequencing (N reads)
chr1
chr1'
chr2
chr2'
chr1
chr1'
Mapping on SCG*
*SCG have 2 copies in recent tetraploid
1Cx = monoploid genome size
1C = holoploid genome size
1C = 2 x 1Cx
Depth
Sequencing depth calculated from 1C = ½ x Sequencing depth calculated from 1Cx
Expected relationship between sequencing depth and depth of mapping on single copy genes
Sequencing depth calculated from 1C
Ploidy=2
Diploids
4
6
8
...
Polyploids
Depth on single copy gene set
Sequencing depth calculated from 1Cx
Diploids
Polyploids
Depth on single copy gene set

FIGURE 1
Schematic representation of the process of mapping reads on single copy genes (SCG) and expected results for a diploid species (top panel) and a tetraploid species (middle panel). When the same number of reads (N) is mapped, the resulting mapping depth on SCGs will be identical for a diploid and a recent tetraploid, provided that all SCGs are still in the duplicated state. The resulting relationship between sequencing depth and depth of mapping on SCG is displayed on the bottom panel: the slope is the same between the diploid and the tetraploid (and any other level of ploidy) when considering 1Cx (right), but decreases with the level of ploidy when considering 1C (left). With no prior knowledge of the ploidy level of the organism sequenced, mapping on SCG genes provides a stable estimate of 1Cx.

Importantly, LocoGSE is a mapping-based estimator that does not rely on a genome assembly. Rather than mapping the reads on a genomic sequence, LocoGSE maps the reads on protein sequences. Short reads are translated into amino-acid sequences and aligned on protein sequences derived from consensus sequences of single copy genes that are shared across all plant lineages.

2.2 Implementation

LocoGSE is coded in Python, runs on a Linux operating system and is included into a Conda environment, since it contains several calls to external programs. It is freely available at https://github.com/institut-de-genomique/LocoGSE.

The program comprises four steps (Supplementary Figure S1). First, sequencing reads are trimmed to 100 nucleotides with Cutadapt (v3.5) (Martin, 2011). In case inputted reads are 100nt long, the users can use the "–no-trim" option to skip the trimming step. Otherwise, the trimming to 100 nt should always be performed: this step is important since the calibration step was performed on 100 nt reads and mapping efficiency is highly dependent on the length of the reads as longer reads are more prone to overlap exon/intron junctions. In a second step, reads are aligned on a set of single copy proteins (by default OneKP ancestral proteins (see section 2.3.3) but the option –busco allows to use BUSCO Embryophyta instead: both protein datasets are provided with the program). The alignment is performed with DIAMOND (Buchfink et al., 2021) (v2.0.14, command "diamond blastx", with e-value parameter set to 0.00001). One best hit per read is then selected and the depth of mapping on each protein is calculated. Subsequently, a filtering step is performed to remove outlier proteins, too highly or too poorly covered compared to the whole single copy protein gene set. Outlier proteins are determined by computing the mapping depth per protein then calculating the Z-Score for each protein and removing the ones with Z-score > 1.96 or< -1.96 (threshold corresponding to P<0.05). These could correspond to genes that have been lost or duplicated in the considered species, or proteins harboring unspecific domains. Finally, the genome size estimation is performed using the following equation, modified from Lander-Waterman (Lander and Waterman, 1988):

$$G = \frac{L}{\beta \text{ x SCP depth}}$$

where L is the cumulative length of the reads used as input for mapping, β is the regression coefficient calculated during the calibration step (default is 1), and SCP depth is the overall depth of mapping of reads on single copy proteins (SCP) after removing the outliers (i.e. total length of reads mapped on all retained SCP divided by the cumulative length of all retained SCP).

A calibration step was performed for Angiosperm lineages in order to calculate β coefficients (see section 2.3). In consequence, for Angiosperm genome size prediction, the user can either provide a plant lineage (listed with the –listlineages option) or a plant family (listed with the –listfamilies option) (Supplementary Figure S2A).

Alternatively, the user can apply LocoGSE to other lineages (animals for instance), and other single copy proteins (BUSCO for instance) and perform their own calibration as explained in the dedicated wiki page: https://github.com/institut-de-genomique/LocoGSE/wiki/2.Linear-regression before providing a slope value with the –slope option (Supplementary Figure S2B).

2.3 Calibration on Angiosperms

Since the genomic reads are mapped on protein sequences (after 6 frame translation), we do not expect to obtain the real sequencing depth when calculating the depth on the single copy protein set. Indeed, reads corresponding to the targeted loci will not be mapped when they happen to fall onto exon/intron or CDS/UTR junctions. Therefore, the method needs to be calibrated using a set of known 1Cx values. The outcome of the calibration is expected to be impacted by the structure of the genes and in particular by the average number of coding exons in the group of species studied.

2.3.1 Retrieving 1C values from Kew db and calculating 1Cx reference values

We extracted prime 1C estimates and associated ploidy levels from Kew Plant DNA C-values database (https://cvalues.science.kew.org/). Estimates with no documented ploidy level in Kew db were not considered. We calculated 1Cx with the following formula:

$$1Cx = 2 * \left(\frac{1C}{ploidy}\right) \text{ (by definition, 1C = 1Cx when ploidy = 2)}$$

When prime estimates for several cytotypes (specimens with different ploidy levels) were available, we discarded species for which there was more than 10% variation among cytotypes in the calculated 1Cx. Among those is the notable example of *Prospero autumnale* for which cytotypes with various chromosome numbers (5, 6, or 7 pairs) and genome sizes have been described (Vestek et al., 2019). For the remaining species, we calculated the mean 1Cx value between cytotypes and considered it as the reference 1Cx for the species.

2.3.2 Angiosperm readsets

We used 430 Angiosperm genome skimming read sets (2 x 100 paired ends illumina reads) from arctic and alpine sampling campaigns (Olofsson et al., 2019; Alsos et al., 2020; Pouchon et al., 2022; Smyčka et al., 2022), for which reference 1Cx genome sizes could be calculated from Kew Plant DNA C-values database (Supplementary Table S1). The samples are broadly distributed among the Monocot and Dicot lineages and their phylogenetic distribution is comparable to that of the species in OneKP (Supplementary Figure S3). Magnolids are absent at the moment, but new calibrations will be performed once more read sets are made public from the PhyloAlps campaign, and include also Gymnosperms, and basal Streptophytes. The sequencing depth of coverage of the 430 read sets varies from 0.017X to 10.13X, with 55% of the samples with depth<1, and 29% between 1 and 2 (Supplementary Figure S4).

2.3.3 Selection of the single copy protein set

We compared two plant single copy gene sets to use for plant genome size estimation in LocoGSE: the widely used BUSCO Embryophyta ancestral proteins (Simão et al., 2015; Manni et al., 2021) (https://busco-data.ezlab.org/v4/data/lineages/embryophyta_odb10.2019-11-20.tar.gz), and the OneKP proteins (Leebens-Mack et al., 2019). OneKP multiprotein alignments were downloaded at https://github.com/smirarab/1kp/blob/master/alignments/alignments-FAA.tar.bz. Consensus sequences were created from multiple alignments using the HMMER3 package version 3.1b1 (Mistry et al., 2013) with the hmmemit function and default parameters.

OneKP contains 1,178 plant transcriptomes (including 658 Monocot and Dicot) that are more diverse phylogenetically than species represented in BUSCO Embryophyta (that contains only 90 Monocot and Dicot genomes, almost half of which are rosids) (Supplementary Figure S3A). Consequently, the consensus derived from OneKP is more representative (in terms of %identity of the matches) of all Angiosperm lineages than BUSCO (Supplementary Figure S3B). Moreover, OneKP contains only 410 ancestral protein sequences, less than half of BUSCO Embryophyta (ODB10) that contains 956 sequences: mapping the reads on OneKP should thus be faster.

Finally, the correlation between the theoretical depth and depths obtained from mapping reads on single copy consensus proteins are slightly better when using OneKP compared to BUSCO (Supplementary Table S2). For all these reasons, we chose to use the OneKP consensus proteins as the default protein database when predicting plant genome sizes with LocoGSE. The databases (OneKP protein consensus sequences and BUSCO Embryophyta) are provided with the program.

2.3.4 Linear regression between depth on OneKP and sequencing depth of coverage

We observe that the depth of mapping on the OneKP single copy gene set (calculated using LocoGSE, Supplementary Figure S2B) and the theoretical sequencing depth calculated from Kew 1Cx are very well correlated (Pearson correlation= 0.89, Pval=2.2e-16, Supplementary Table S2). In addition, their relationship is close to linear (Supplementary Figure S5A). We noticed that P-values for Pearson correlations and R2 of the regressions were slightly higher when separating plant lineages (Supplementary Table S2) rather than considering all samples together. Thus, we performed separate calibrations for each lineage. Our approach uses a linear regression to calculate the coefficient (β) by which to multiply the depth on OneKP single copy proteins in order to obtain the sequencing depth with the following formula:

$$Estimated\ \ 1Cx\ \ depth = \ \beta \text{ x OneKP depth}$$

For each group, we estimated the β coefficient with a robust linear regression (Figure 2). Robust linear regressions were performed with the Rpackage *robust* (*lmrob* function) and allowed to minimize the influence of outliers on the results of the regression. The β coefficients vary from 1.11 to 1.895 across lineages, and the coefficient obtained on all lineages together is of 1.56 (Supplementary Table S2), and could be used in the absence of information on the plant lineage.

In order to check the rationale of our approach, we also calculated the Pearson correlation between the depth on OneKP and the sequencing depth calculated from Kew 1C: as expected, the correlation coefficient (0.83) is lower than that calculated from 1Cx. Moreover, the samples behave differently, according to their degree of ploidy (Supplementary Figure S5B). Therefore, we confirm that, as other methods, LocoGSE is indeed a monoploid genome size (1Cx) estimator.

2.4 Benchmark (data and programs)

2.4.1 Readsets

In order to benchmark LocoGSE, we downloaded short reads datasets for 8 Angiosperm species broadly distributed across the phylogeny and with genome sizes (1Cx) ranging from 157 Mb to 17.5 Gb and sequencing depths (calculated from Kew 1Cx) ranging from 0.64X to 60.1X (Supplementary Table S3). We also downloaded genome assemblies in order to test mapping based approaches, when available (all species except *Papaver nudicaule*). Among the read sets selected, some were also included in the benchmark performed by Pucker et al. for MGSE: *V. vinifera* and *Z. mays* (Pucker, 2019). One sample is highly heterozygous (*Vitis vinifera*) and one is tetraploid (*Solanum tuberosum*). For *Vitis vinifera*, since the readset was very large (380X), we randomly selected 10% of the read pairs to perform our benchmark. For *Solanum tuberosum* and *Zea mays*, we downloaded only one of the two files, containing one read from each pair. For *Allium cepa*, we downloaded two read sets and ran the prediction on each one independently, in order to check the consistency of the results (Supplementary Table S4).

2.4.2 Programs

We compared LocoGSE with five short reads-based genome size estimators: three based on k-mers [GenomeScope2.0 (Ranallo-Benavidez et al., 2020), CovEst (Hozza et al., 2015) and RESPECT (Sarmashghi et al., 2021)] and two based on mapping [ModEst (Pfenninger et al., 2022) and MGSE (Pucker, 2019)].

Before running GenomeScope and CovEst, we first generated the k-mer spectrums using Jellyfish (Marçais and Kingsford, 2011), either with k-mers of 21 or 31. The CPU time and memory usage of Jellyfish were included in the performances measured for GenomeScope and CovEst.

GenomeScope (version 2.0) was downloaded at https://github.com/tbenavi1/genomescope2.0. We set up the parameters to k=31 since it is a broadly used k-mer size and used default ploidy (p=2): the output "genome haploid length" corresponds to 1Cx.

CovEst was downloaded at https://github.com/mhozza/covest and launched with k=21 (default) or k=31, and with -m repeats since plant genomes are known to be rich in repeated elements (as expected, tests with the default option (-m basic) produced less accurate genome size estimations).

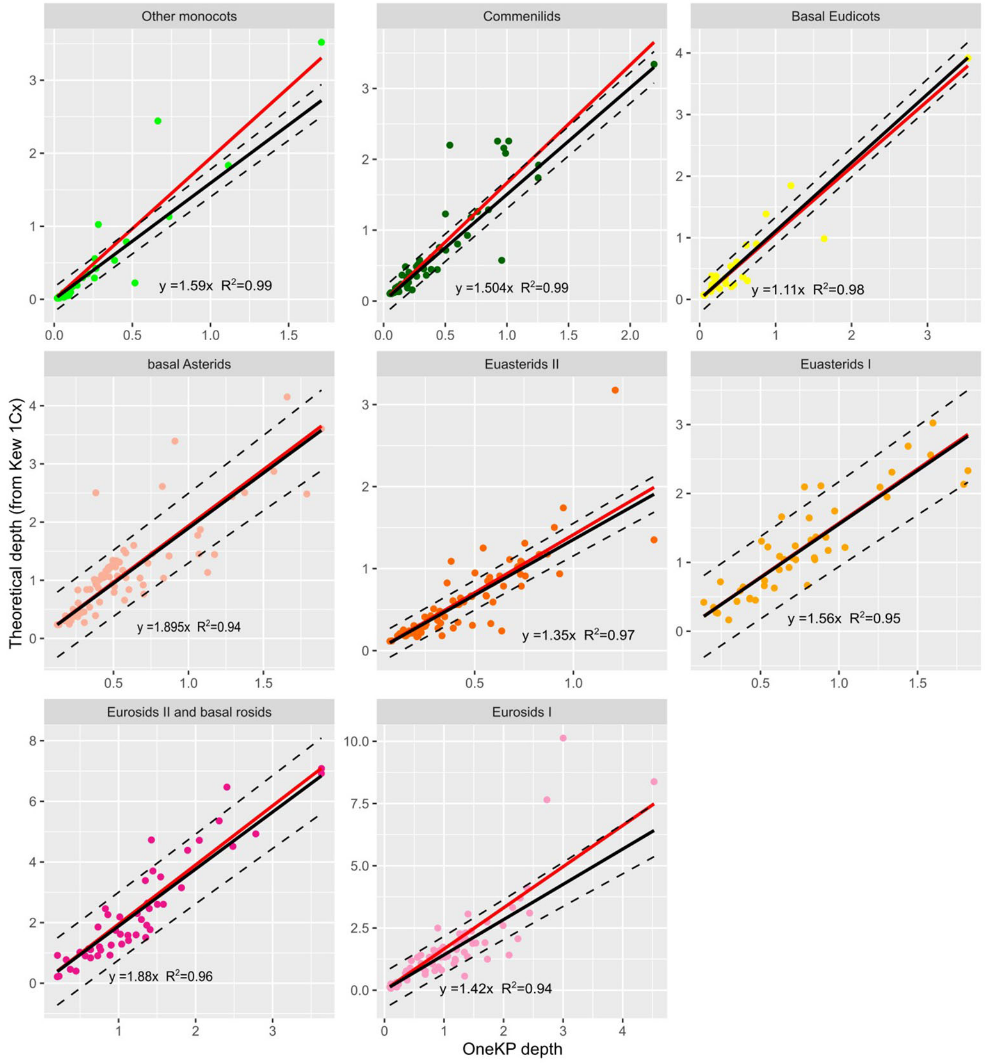

FIGURE 2
Relationship between depth on OneKP single copy genes and theoretical depth (calculated from Kew 1Cx) in the training set, for 8 plant lineages. Black line is the regression line obtained after robust regression, red line is the regression line obtained with standard regression. Regression line equations and R coefficients are displayed for each lineage.

RESPECT was downloaded from https://github.com/shahab-sarmashghi/RESPECT. It is to be noted that RESPECT requires the installation of an academic licence for Gurobi, and thus can not be run on any computer, which is a limitation. RESPECT was launched with default parameters, on all reads and also on reads filtered with Kraken2 (Wood et al., 2019) to remove human and bacterial contamination according to the authors' recommendations. We used the standard database downloaded from https://genome-idx.s3.amazonaws.com/kraken/k2_standard_20230314.tar.gz. Kraken unclassified reads (option –unclassified-out) were provided as input for RESPECT. Genome size estimations obtained with or without filtering the reads were very similar, which suggests that the datasets were not very contaminated. The CPU time consumption was lower on the filtered reads, but when adding the time necessary to run Kraken2, the performances were comparable (Supplementary Table S4).

Before running mapping based predictors, we first aligned the reads onto the assemblies using BWA-MEM (Li, 2013). MGSE was downloaded from https://github.com/bpucker/MGSE. It was launched using BUSCO annotations, which required first running BUSCO (Simão et al., 2015) on the assembly, with options "–augustus –lineage embryophyta_odb10". We then provided MGSE with the BUSCO output directory (–busco option).

ModEst was downloaded from https://github.com/schellt/backmap and launched with default parameters. In one case (*Z. mays* readset), ModEst failed and provided a fake depth of "1", leading to an aberrant size estimation instead of generating an error message. After relaunching several times, the same error occurred.

For each program, we monitored the memory consumption and the CPU time (user+system) (Supplementary Table S4). For each genome size prediction, we also calculated the % of error as:

$$(Predicted\ Size\ -\ Expected\ Size\ Kew)\ x\ 100/Expected\ Size\ Kew$$

2.5 Genome size predictions on read sets at various sequencing depths

In order to evaluate the sequencing depth required for the genome size predictors, we built datasets with various depths of coverage for *Chenopodium suecicum* by randomly selecting read pairs in the fastq files (SRR4425238, total= 28.9X) (Supplementary Table S5). Subsampling was performed using an in-house program, getRandomSeq, available at https://github.com/institut-de-genomique/saturn. For each depth, we ran LocoGSE, GenomeScope 2.0 (k=31), RESPECT and CovEst (with k=31 since the prediction was more accurate than for default value of k=21 on this dataset), and computed the average and standard deviation for the predictions as well as the % of error of the predictions, by comparison to the 1Cx reference value obtained from Kew db.

2.6 Ploidy estimation

We downloaded 12 read sets from *Senecio doronicum* specimens with various degrees of ploidy (Supplementary Table S6) and for which the 1C genome size was estimated by flux cytometry (Fernández et al., 2022). We used LocoGSE to estimate 1Cx genome sizes, and we estimated the ploidy level of each sample with the following formula:

$$Estimated\ \ ploidy = 2\ * \frac{1C\ \ (cytometry)}{1Cx\ \ (LocoGSE)}$$

3 Results

3.1 Genome size predictions obtained on 430 Angiosperm species (used for calibration)

As a first check for the validity of the approach, we compared genome sizes predicted by LocoGSE with expected 1Cx genome sizes from Kew database on the 430 plant samples used for the calibration of the method (Figure 3). For most lineages, the predictions were accurate (mean error rates are between 0.21 for Euasterids II and 0.37 for other monocots). However, sizes were underestimated for very large genomes (in particular “other monocots”: Supplementary Figure S6). This observation can be explained by the fact that genome skimming experiments provide a very low depth of coverage (Supplementary Figure S7) for such large genomes, and the estimation of mapping depth on single copy proteins with less than 1 read per gene becomes very uncertain. As it is not surprising that the results obtained on the training set used to calibrate the method are accurate, we benchmarked LocoGSE and other genome size estimators on independent datasets.

3.2 Comparison of LocoGSE with other genome size estimators

3.2.1 Prediction accuracy

We downloaded read sets that were not used for the calibration and correspond to eight species from all major plant lineages with various expected sizes and various sequencing depths (Supplementary Table S3). We compared LocoGSE predictions with five other genome size prediction softwares (two k-mer based and two assembly mapping-based) (Supplementary Table S4, Supplementary Figures S4, S8). For CovEst, we tested two values of k-mer (the default k=21, and the more commonly used k=31): the results were similar, and the best prediction was alternatively the one with k=21 or k=31. We tested RESPECT with and without removing contaminant reads with Kraken (see Methods): again, results were very similar. For subsequent comparisons, we focused on default options for all programs (Table 1). Kew 1Cx estimates were used as the reference for genome size because complete (T2T) assemblies were not available for all plant lineages that were included in the benchmark. Moreover, the added-value of Kew genome sizes is that they are obtained by an orthogonal method (i.e. flow cytometry, that is not sequence-based) compared to the sequence-based genome size estimators that are benchmarked. As a matter of comparison, Table 1 also displays the size and level of the most complete assembly for each species. At very low depth of coverage (*P. nudicaule* 0.6X and *A. cepa* 3X), LocoGSE and RESPECT were the only softwares to provide non aberrant genome sizes. At low depth (*H. annuus* 10X), all softwares except GenomeScope provided acceptable results, but the best prediction (with regard to Kew estimate) is the one provided by LocoGSE. When considering the chromosome level assembly size as reference, four predictors including LocoGSE provide very accurate predictions (error rate<0.1). Above 25X, all programs are able to make acceptable predictions, but some produce aberrant predictions for some read sets (ModEst for *Z. mays*, which is probably due to the error reported in Methods, RESPECT for *Z. mays*, *A. thaliana* and *S. tuberosum*, which is not surprising since it was designed for low coverage samples) (Supplementary Figure S8). For *Vitis vinifera*, which is a highly heterozygous sample, the predictions that were within an acceptable range of the expected size were the ones provided by LocoGSE and GenomeScope. For *Solanum tuberosum*, which is a tetraploid, LocoGSE, GenomeScope and the two mapping-based approaches provided predictions that are close to 1Cx, strongly supporting our observation that such approaches are expected to estimate 1Cx rather than 1C (Figure 1). In summary, predictions made by LocoGSE are never aberrant (their %error range is the lowest of all predictors: Figure 4), and often the best ones (Table 1, Supplementary Figure S8). It is to be noted that few genome size predictions are below the threshold of 10% of error, underlining the difficulty of sequence-based genome size estimation.

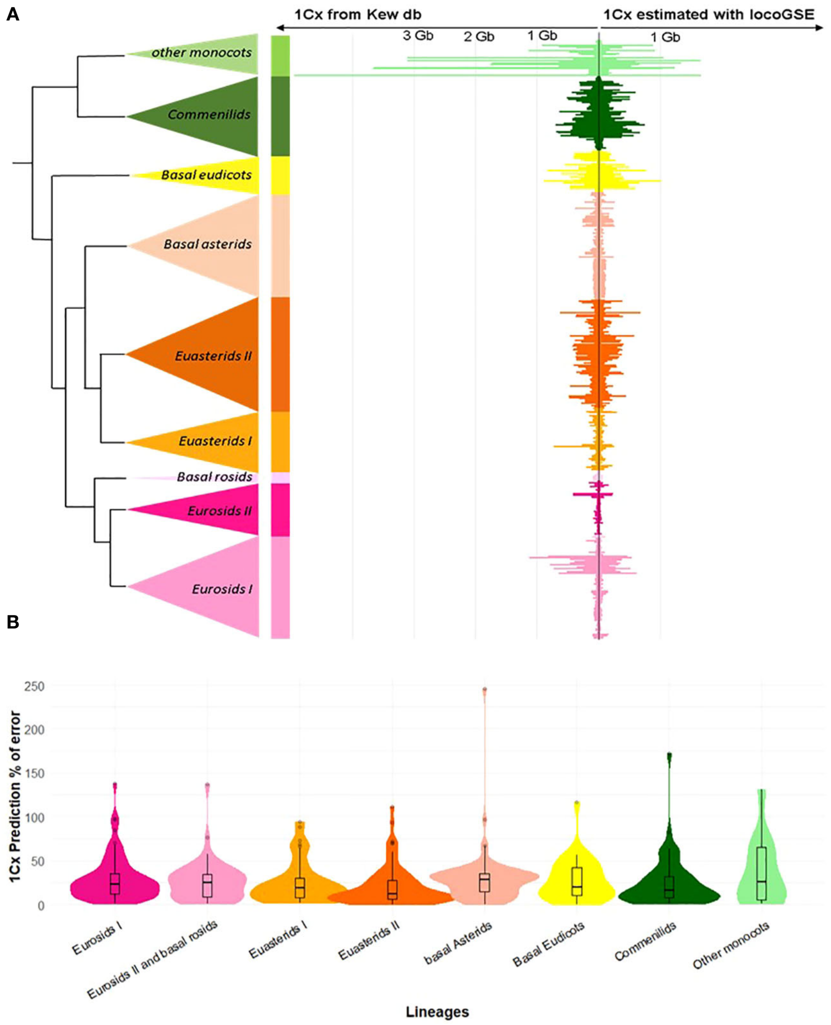

FIGURE 3
Results of genome size estimation with LocoGSE on the training set. **(A)**. Comparison of expected (left: Kew 1Cx sizes) and predicted sizes (right: LocoGSE output) for the 430 species of the training set, ordered by phylogeny. **(B)**. Violin plots representing the % of error of LocoGSE predictions by plant lineage. Median % of errors for each lineage are displayed above plots.

3.2.2 Performances

We compared CPU run time and memory consumption for the six programs with default parameters (Supplementary Table S4). Figure 5 displays the program performances, with the datasets (species on the x axis) ordered from the lowest to the highest number of nucleotides in the inputted readsets. As expected, for all programs, the running time increases when increasing the number of nucleotides (Figure 5A). Additionally, the CPU time is one order of magnitude higher for the two assembly mapping-based methods (this is caused by the step of mapping the reads on the assembly). LocoGSE requires only to map the reads on a few hundred of single copy proteins, which explains why it is faster. As expected, k-mer based approaches are the least time-consuming. However the algorithm used in RESPECT relies on a complex modeling of genomic parameters (Sarmashghi et al., 2021) and thus performs slower. In summary, the CPU time for LocoGSE is comparable to that of RESPECT and intermediary between k-mer and assembly mapping-based methods. As CPU time, memory usage also increases with the number of nucleotides treated but LocoGSE is a notable exception: memory usage remains constant and low for all runs (Figure 5B).

3.2.3 Required sequencing depth

In order to identify the short read sequencing depth required for each software to provide accurate predictions, we compared the genome size estimations obtained from read sets sampled from the same sequencing run at various depths (Figure 6). We focused on the three k-mer based programs and LocoGSE, since the need for an assembly precludes the use of MGSE and ModEst on low coverage datasets. We selected the readset from *Chenopodium suecicum*, because all estimators provided accurate genome size predictions when using the whole dataset (28.9X) (Supplementary Figure S8). First, it is notable that predictions are usually very consistent between samples (Supplementary Table S5). Nonetheless, as expected, LocoGSE predictions are more variable between samples (larger error bars) at extremely low depth of coverage (<0.5X) than at higher depths. Moreover, all predictions appear to

TABLE 1 Genome size estimations obtained with 6 predictors applied to 8 plant datasets with default parameters. More extensive information can be found in Supplementary Table S4. Estimators are ranked according to the error rate (absolute value), calculated as explained in Methods by comparison with 1Cx obtained from Kew db and assembly size (when available). All genome sizes are provided in Megabases.

Species	*Papaver nudicaule*			*Allium cepa*			*Helianthus annuus*			*Zea Mays*			*Chenopodium suecicum*			*Arabidopsis thaliana* col-0			*Vitis vinifera* (Chardonnay)			*Solanum tuberosum*		
Lineage	Basal Eudicots			Other monocots			Euasterids I			Commenilids			Basal Asterids			Eurosids II			Eurosids I			Euasterids II		
SRA_ID	SRR17698145			ERR5262394			SRR5004592			SRR1575500			SRR4425238			SRR1810274			SRR7141304			SRR15198297		
sequencing depth (after trimming to 100nt)	0.65 (0.43)			3.44 (2.73)			9.60 (9.60)			25.61 (25.61)			28.87 (28.87)			37.71 (37.71)			38.03 (25.46)			60.10 (39.80)		
1Cx from Kew (Mb)	4018			17542			3597			2646			739			157			392			948		
Most complete assembly length (Mb)	No assembly			15932+			3010+			2288+			537			142++			490			775+		
GS LocoGSE*	2338	1	-	11836	2	2	3212	1	**4****	1859	1	1	688	**4****	3	204	4	4	481	1	**2****	1003	**1****	4
GS GenomeScope 2.0 (k=31)*	10	4	-	205	5	5	425	6	6	1366	4	4	719	**1****	4	158	1**	2	488	2	**1****	780	3	**1****
GS CovEst (k=21)*	355	3	-	10	6	6	3183	3	**2****	1480	3	3	515	6	**1****	168	**2****	3	907	5	5	1289	5	5
GS RESPECT*	1532	2	-	18844	**1****	1	3199	2	**3****	9384	5	5	781	**2****	5	3176	6	6	168	3	4	85665	6	6
GS ModEst*	No assembly	-	-	28050	3	3	3070	4	**1****	64500	6	6	783	**3****	6	310	5	5	1020	6	6	887	**2****	3
GS MGSE (BUSCO)*	No assembly	-	-	2800	4	4	2487	5	5	1728	2	2	671	**5****	2	135	3	**1****	713	4	3	780	3	**1****

+:assemly level=chromosome.

++:assemly level=complete.

*:2nd col=rank of the estimator with ref=Kew, 3rd col= rank of the estimator with ref=Assembly.

**:less than 10% error (bold).

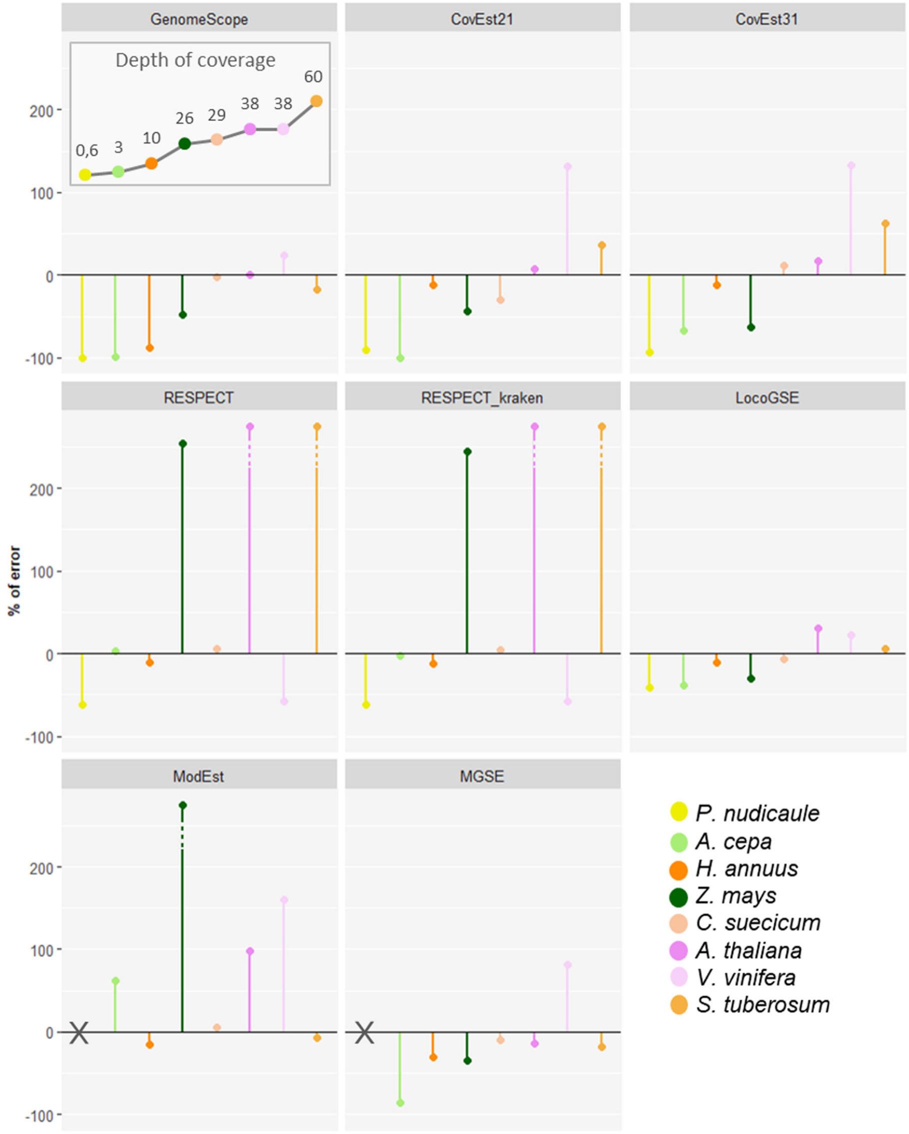

FIGURE 4
Percent of error (relative to Kew 1Cx) obtained for genome size predictions for 6 genome size estimators on 8 plant sequencing datasets, with various sequencing depths of coverage. Plant species are ordered from left to right from the lowest to the highest sequencing depth as displayed in the "depth of coverage" panel: points are connected for easier visualization.

reach a plateau after a certain depth. Interestingly, the plateau is reached at low depths for LocoGSE and RESPECT, whereas it is reached at 10X for CovEst, and 26X for GenomeScope (Figure 6A). When comparing the two packages designed for very low coverage datasets, we notice that both LocoGSE and RESPECT reach a plateau at very low depths (0.5X). Notably, LocoGSE provides better predictions at extremely low depth (0.1X) and converges faster than RESPECT towards accurate estimations (Figures 6B, C). Both tools are very accurate on the complete *C. suecicum* dataset but it remains to be noted that RESPECT is not recommended for high coverage datasets (Sarmashghi et al., 2021), and was shown to produce erroneous estimations on other read sets at high coverage. In particular, for the *Arabidopsis thaliana* dataset, with a depth of coverage of 38X, RESPECT estimated a genome size of 3,176 Mb instead of the expected 157 Mb (Table 1).

In conclusion, LocoGSE is the only software that can be used at any sequencing depth, and although the running time is higher for LocoGSE compared to k-mer based approaches like GenomeScope for a given readset, the number of reads required is much lower (1X vs >25X). Consequently, the running time needed to get a reasonable prediction is actually lower for LocoGSE than GenomeScope.

3.3 Comparison of predictions for various degrees of ploidy

We wanted to investigate the sensitivity of the method to the ploidy level of the input samples. For that purpose, we used various read sets from *Senecio doronicum* specimens, with ploidy levels ranging from 4 to 8 (Fernández et al., 2022) (Supplementary Table S6). As seen

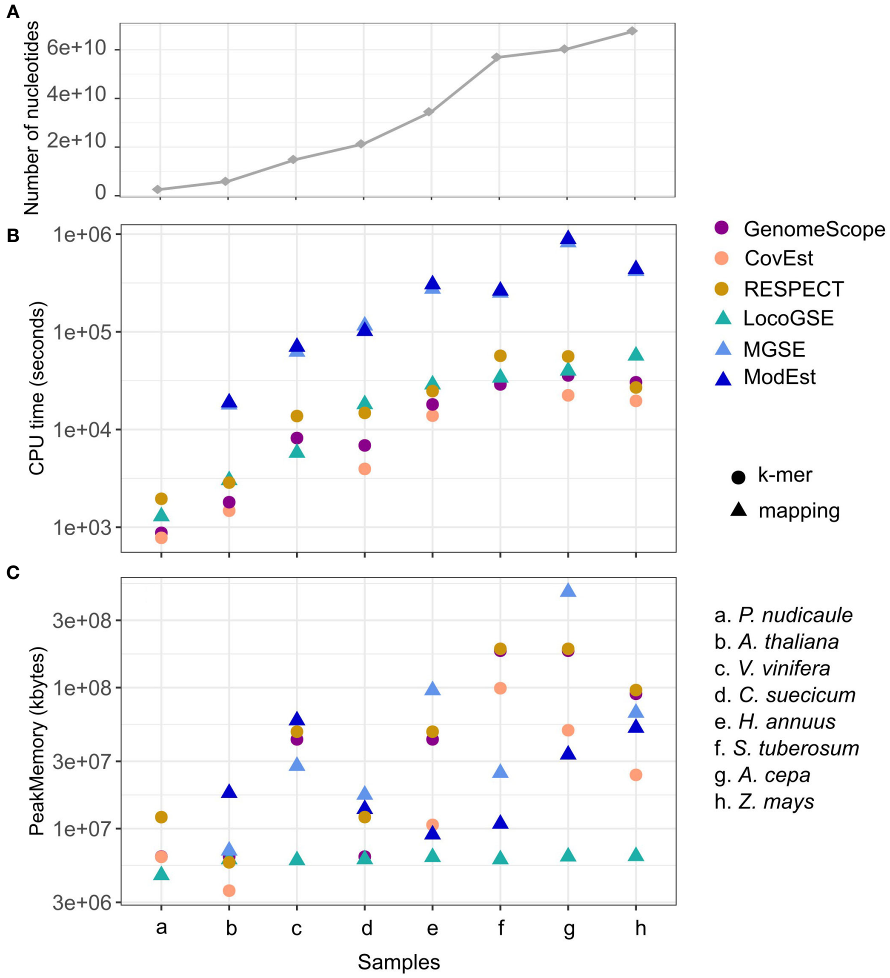

FIGURE 5
Performances of 6 genome size estimators for 8 plant short read datasets. **(A)** number of nucleotides in the sample. **(B)** CPU time (log scale) and **(C)** Peak Memory. Datasets are ordered from the lowest to the highest number of nucleotides treated (the number of reads and their lengths can be found in Supplementary Table S3), and points in panel **(A)** are connected for easier visualisation. K-mer-based methods are represented with circles, and mapping-based methods with triangles.

in Figure 7, 1Cx genome sizes estimated with LocoGSE are close to expectations and stable across all samples, regardless of the ploidy level. We can thus conclude that LocoGSE is effectively estimating 1Cx rather than 1C, as expected from our model (Figure 1). We estimated the ploidy level using 1C genome size measured by flow cytometry and 1Cx genome size estimated by LocoGSE. Estimated ploidies are close to observed ones. Notably, for higher degrees of ploidy, the ploidy levels tend to be slightly underestimated as a consequence of lower 1C estimation by flow cytometry but constant 1Cx estimation by LocoGSE. This result could be attributed to genome downsizing after polyploidy, that probably did not affect single copy OneKP genes as much as the rest of the genome, thus resulting in a slight overestimation of the 1Cx genome size by LocoGSE.

4 Discussion

We developed LocoGSE, a monoploid genome size estimator aimed at using short reads at very low coverage. Although it is not often stated in the documentation of other genome size estimators, all sequence-based genome size estimators (k-mer or mapping based) predict 1Cx rather than 1C, since they are not able to distinguish k-mers or reads from different haplotypes under a certain threshold of heterozygosity. GenomeScope 2.0 (Ranallo-Benavidez et al., 2020), the most widely used k-mer based method, is able to cope with highly heterozygous and polyploid samples. The so-called “genome haploid length” it predicts indeed corresponds to 1Cx. But it is to be noted that GenomeScope needs the user to provide a ploidy value in order to estimate 1Cx accurately. Otherwise, it is estimated with the default value of 2. Since most lineages are diploids, the absence of ploidy information is not often problematic: the estimated 1Cx value is equal to 1C in this case. We believe this is the reason why “genome size estimators” do not document the fact that the estimation they provide corresponds to 1Cx (monoploid genome size). But for lineages where polyploidy is frequent, like plants (Heslop-Harrison et al., 2023), genome size predictors will need to be combined with short-reads based reference-free ploidy estimators such as SMUDGEPLOT

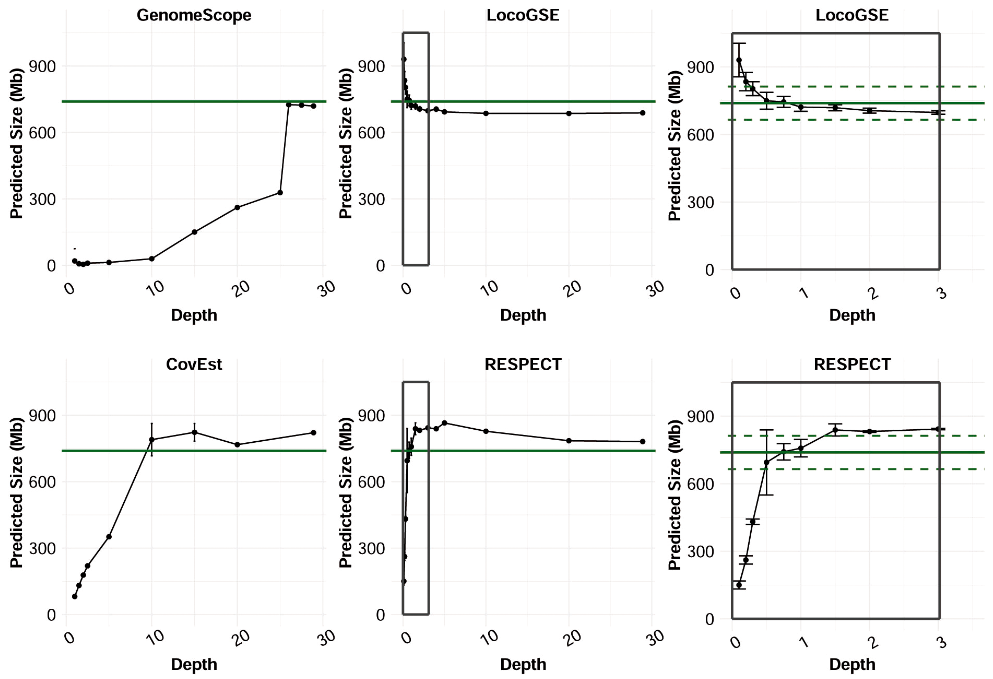

FIGURE 6
Genome size estimations made by predictors on subsets of reads of *C. suecicum* with increasing depth of coverage, from 0 to ~30X, and zoom on 0 to 3X for LocoGSE and RESPECT. Several runs (up to 10) were performed for each depth: dots are positioned at the average prediction and error bars represent one standard deviation up and one standard deviation down. Green continuous line shows the expected genome size (Kew 1Cx estimate: 739 Mb) and dashed lines the 10% error interval.

(Ranallo-Benavidez et al., 2020) or PLOIDYFROST (Sun et al., 2023) in order to obtain an estimation of 1C genome size. In this aim, there is a need to develop ploidy predictors that are able to cope with low coverage datasets. Nevertheless, monoploid genome size (1Cx) predictions also have inherent value: they can be used to estimate the ploidy level of a specimen, provided flow cytometry is performed to estimate 1C. Such an approach could be easier to implement than classical microscopy approaches to infer the number of chromosomes and be of great use for botanists.

LocoGSE relies on deducing the sequencing depth from the depth of mapping of short reads on single copy genes. The fact that the coefficient (slope) linking depth of mapping on the single copy genes and sequencing depth is not equal to 1 could seem counter-intuitive. In the model described by Pflug et al., the coefficient is assumed to be 1, which is expected when mapping the reads on complete gene sequences, at the DNA level (Pflug et al., 2020). Contrastingly, for LocoGSE, translated short DNA reads are mapped on protein sequences. Consequently, various possible biases are likely to impact the quality of mapping (and subsequent depth calculation) and explain why different coefficients are found for different phylogenetic branches. First, the phylogenetic distance between the sample studied and the consensus for monocopy genes will have an impact on the number of reads mapped. However, we showed that the OneKP consensus, unlike BUSCO, aligns with similar percent identities on all Angiosperm lineages (Supplementary Figure S3), suggesting it is evenly distant to all lineages, which is the reason why we selected it as the default dataset. Additionally, for lineages with different numbers of exons/introns per gene, the results of the mapping will be different, since reads will not map on exon-exon junctions. Since OneKP genes are ancestral to all *Viridiplantae*, intron/exon structures of these genes are expected to be relatively stable, but in the event of species harboring very small exons (shorter than 100nt, the length or trimmed reads mapped), we expect LocoGSE not to be able to predict genome size accurately. Also, for polyploids that are in the process of diploidization by loss of copies of genes to go back to the diploid state (Langham et al., 2004), the signal can be noisy because some of the supposed "single copy" genes are in several copies and others in single copy, leading to a predicted genome size between 1Cx and 1C. Finally, LocoGSE is expected to be sensitive to contaminations in the read sets (bacterial DNA, chloroplastic DNA), because it is relying on the Lander-Waterman equation, and even more so when it is used on low coverage readsets. For this reason, we filtered out two genome skimming read sets from the training set (used for calibration), because they contained more than 20% of bacterial reads (all others had less than 6% of bacterial reads). We used Kraken2 on the read sets used for benchmarking the 6 genome size estimators and showed that at most 7% of the reads correspond to bacterial or human contamination. Such low percentages allow accurate gene prediction, but one should keep in mind that it is recommended to check the level of contamination of the read sets before running any sequence-based genome size estimator.

As already mentioned, LocoGSE requires a calibration step using short reads datasets corresponding to genomes with known 1Cx sizes. This was achieved using a large dataset of plant genome skims as well as carefully curated genome size prime estimates from Kew Plant DNA C-values database. While flow cytometry may not always be

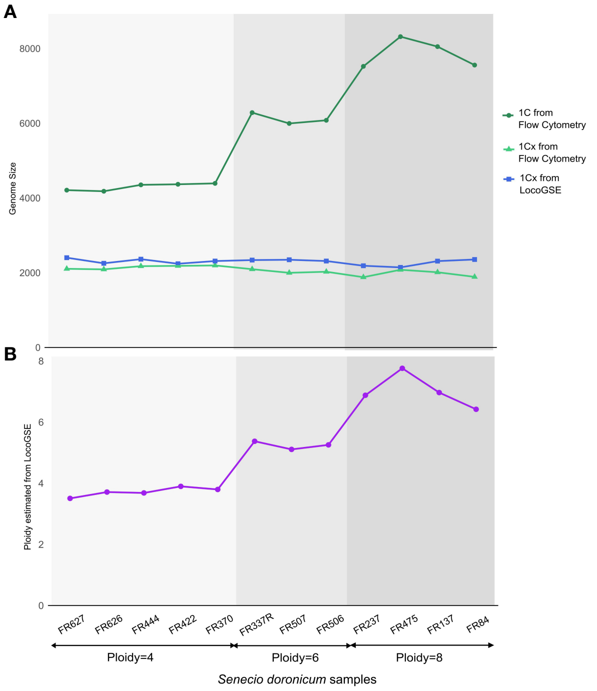

FIGURE 7
Genome size and ploidy estimations for 12 *Senecio doronicum* specimens with various ploidy levels (4, 6, 8). **(A)** genome size estimated by flow cytometry (blue) and LocoGSE (green). **(B)** ploidy estimated by dividing Kew 1C by LocoGSE 1Cx (purple). Samples are ordered according to their ploidy levels and points are connected for easier visualization.

entirely accurate, Kew size estimates stand out as the closest approximation to the ground truth that we have at our disposal for a very large number of species. Moreover, we carefully selected the species to include in the calibration set among the ones with low 1Cx variation among cytotypes, in order to overcome the effect of intraspecies variation of monoploid genome size. Indeed, intraspecies genome size variation has been described in various plants (Schmuths et al., 2004; Díez et al., 2013; Long et al., 2013; Bilinski et al., 2018; Becher et al., 2021; Balant et al., 2022) and genome size should be considered as a trait of individuals rather than species. Consequently, users should be aware that size estimations generated by any method correspond to the analyzed individual and are not necessarily representative of the whole species.

Using LocoGSE on plants is very straightforward, because of the added-value provided by the calibration we performed. For other lineages, the user will need to input a set of single copy genes corresponding to the organism studied. BUSCO database provides gene sets for a wide range of species (Simão et al., 2015; Manni et al., 2021). Alternatively, one can build orthogroups for a specific lineage, detect single copy gene families, and generate consensus sequences for these families, as was done recently to develop a single copy genes resource for coral genomes (Noel et al., 2023). Then, if possible and instead of using a default slope of 1, the user can calibrate LocoGSE using sequencing data from genomes of known sizes, as explained in our wiki (https://github.com/institut-de-genomique/LocoGSE/wiki) (Supplementary Figure S2B).

To summarize, we showed that LocoGSE fills a gap in sequence-based genome size estimation for several reasons. First, the depth of coverage required to get optimal predictions is as low as 0.5X, and no reference assembly is required, which makes LocoGSE particularly suitable for genome skimming datasets. LocoGSE can also find its use in biodiversity sequencing projects in which an inexpensive and superficial first sequencing experiment can provide a lot of information including genome size. Another package, RESPECT (Sarmashghi et al., 2021), also performs very well in the same range of depth. However, its use is complicated by the need to install an academic licence (many research institutions, although public and non-profit, do not have the “academic” status). More worryingly, RESPECT sometimes provides aberrant results when used on higher coverage readsets, which precludes its use on unknown genomes with no previously reported genome sizes. On the contrary, although LocoGSE was calibrated on very low depth of coverage datasets (<2X) and is thus expected to work better for low depths, we showed that the estimations made at higher depths are also accurate. At high coverages (above 30X), users may choose to use k-mer based approaches like GenomeScope, that are less time consuming and were shown to provide accurate results. However, another possible option would be to apply LocoGSE on a subset of reads, which would be more memory-efficient and probably even faster.

Second, as LocoGSE relies on mapping on consensus (ancestral) sequences of single copy genes, it is not sensitive to possible heterozygosity of the considered specimen. Here again, no prior knowledge about the level of heterozygosity or the ploidy of the sample treated is required. Consequently, since LocoGSE can be launched on low coverage sequencing data without any prior knowledge about the properties of the genome considered (size, ploidy, heterozygosity), it can be of great use to inform on sequencing strategies for environmental samples.

Finally, thanks to the calibration already performed, the current version of LocoGSE is readily applicable to most Angiosperm lineages (pre-computed slopes for OneKP (default) and BUSCO Embryophyta single copy gene sets). We will update the dataset with more plants and extend the calibration to Gymnosperms and other vascular plants in the near future.

In conclusion, we believe that LocoGSE will be of great use for the community of researchers in the fields of environmental genomics and plant genome size analysis.

5 Collaborators

5.1 Members of the PhyloNorway consortium

I.G. Alsos, M.K. Føreid Merkel, Y. Lammers (The Arctic University Museum of Norway, Tromsø, NO), E. Coissac, C. Pouchon (Laboratoire d’Ecologie Alpine, CNRS, UGA, Grenoble, FR); A. Alberti, F. Denoeud, P. Wincker (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris- Saclay, FR).

5.2 Members of the PhyloAlps consortium

S. Lavergne, C. Pouchon, E. Coissac, C. Roquet, J. Smyčka, M. Boleda, W. Thuiller, L. Gielly, P. Taberlet, D. Rioux, F. Boyer, A. Hombiat, B. Bzeznik (Laboratoire d'Ecologie Alpine, CNRS, UGA, Grenoble, FR); A. Alberti, F. Denoeud, P. Wincker, C. Orvain (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Univ. Paris-Saclay, FR); C. Perrier, R. Douzet, M. Rome, J.G. Valay, S. Aubert (Jardin Alpin du Lautaret, CNRS, UGA, Grenoble, FR); N. Zimmermann, R. O. Wüest, S. Latzin, S. Wipf (Swiss Federal Research Institute WSL, Birmensdorf, CH); J. Van Es, L. Garraud, J.C. Villaret, S. Abdulhak, V. Bonnet, S. Huc, N. Fort, T. Legland, T. Sanz, G. Pache, A. Mikolajczak (Conservatoire Botanique National Alpin, Gap, FR); V. Noble, H. Michaud, B. Offerhaus, M. Pires, Y. Morvant (Conservatoire Botanique National Méditerranéen, Hyères, FR); C. Dentant, P. Salomez, R. Bonet (Parc National des Ecrins, Gap, FR); T. Delahaye (Parc National de la Vanoise, Chambery, FR); M.F. Leccia, M. Perfus (Parc National du Mercantour, Nice, FR); S. Eggenberg, A. Möhl (Info-Flora, Bern, CH); B. Hurdu, M. Puşcaş (Babeş Bolyai University, Institute of Biological Research, Cluj Napoca, RO), M. Slovák (Institute of Botany, Bratislava, SK).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author. The program is available at https://github.com/institut-de-genomique/LocoGSE.

Author contributions

PG-T: Writing – review & editing, Writing – original draft, Software, Methodology. BI: Writing – review & editing, Software, Methodology. IA: Writing – review & editing, Resources. EC: Writing – review & editing, Resources. SL: Writing – review & editing, Resources. J-MA: Writing – review & editing, Validation, Supervision. FD: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Genoscope and the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA). Sequencing was performed within the framework of the PhyloAlps project, funded by France Génomique (ANR-10-INBS-09-08) and the PhyloNorway project funded by the Research Council of Norway (226134/F50) and the Norwegian Biodiversity Information Centre (14-14, 70184209).

Acknowledgments

We would like to thank Dr Ilia Leitch for providing us with curated Kew genome sizes that were not all available at the beginning of the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1328966/full#supplementary-material

References

Alsos, I. G., Lavergne, S., Merkel, M. K. F., Boleda, M., Lammers, Y., Alberti, A., et al. (2020). The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants Basel Switz.* 9, 432. doi: 10.3390/plants9040432

Balant, M., Rodríguez González, R., Garcia, S., Garnatje, T., Pellicer, J., Vallès, J., et al. (2022). Novel Insights into the Nature of Intraspecific Genome Size Diversity in Cannabis sativa L. *Plants Basel Switz.* 11, 2736. doi: 10.3390/plants11202736

Becher, H., Powell, R. F., Brown, M. R., Metherell, C., Pellicer, J., Leitch, I. J., et al. (2021). The nature of intraspecific and interspecific genome size variation in taxonomically complex eyebrights. *Ann. Bot.* 128, 639–651. doi: 10.1093/aob/mcab102

Bennetzen, J. L., MA, J., and Devos, K. M. (2005). Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95, 127–132. doi: 10.1093/aob/mci008

Bilinski, P., Albert, P. S., Berg, J. J., Birchler, J. A., Grote, M. N., Lorant, A., et al. (2018). Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. *PLoS Genet.* 14, e1007162. doi: 10.1371/journal.pgen.1007162

Blommaert, J. (2020). Genome size evolution: towards new model systems for old questions. *Proc. R. Soc B Biol. Sci.* 287, 20201441. doi: 10.1098/rspb.2020.1441

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* 34, 247–278. doi: 10.1242/jcs.34.1.247

Chase, M. W., Samuel, R., Leitch, A. R., Guignard, M. S., Conran, J. G., Nollet, F., et al. (2023). Down, then up: non-parallel genome size changes and a descending chromosome series in a recent radiation of the Australian allotetraploid plant species, Nicotiana section Suaveolentes (Solanaceae). *Ann. Bot.* 131, 123–142. doi: 10.1093/aob/mcac006

Chen, S. H., Rossetto, M., van der Merwe, M., Lu-Irving, P., Yap, J.-Y. S., Sauquet, H., et al. (2022). Chromosome-level *de novo* genome assembly of Telopea speciosissima (New South Wales waratah) using long-reads, linked-reads and Hi-C. *Mol. Ecol. Resour.* 22, 1836–1854. doi: 10.1111/1755-0998.13574

Coissac, E., Hollingsworth, P. M., Lavergne, S., and Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* 25, 1423–1428. doi: 10.1111/mec.13549

Dai, S.-F., Zhu, X.-G., Hutang, G.-R., Li, J.-Y., Tian, J.-Q., Jiang, X.-H., et al. (2022). Genome size variation and evolution driven by transposable elements in the genus oryza. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.921937

Díez, C. M., Gaut, B. S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L. E., et al. (2013). Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytol.* 199, 264–276. doi: 10.1111/nph.12247

Dolezel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244. doi: 10.1038/nprot.2007.310

Fernández, P., Hidalgo, O., Juan, A., Leitch, I. J., Leitch, A. R., Palazzesi, L., et al. (2022). Genome Insights into Autopolyploid Evolution: A Case Study in Senecio doronicum (Asteraceae) from the Southern Alps. *Plants Basel Switz.* 11, 1235. doi: 10.3390/plants11091235

Fu, C.-N., Mo, Z.-Q., Yang, J.-B., Cai, J., Ye, L.-J., Zou, J.-Y., et al. (2022). Testing genome skimming for species discrimination in the large and taxonomically difficult genus Rhododendron. *Mol. Ecol. Resour.* 22, 404–414. doi: 10.1111/1755-0998.13479

Gilbert, D. G. (2022). Genes ruler for genomes, Gnodes, measures assembly accuracy in animals and plants. doi: 10.1101/2022.05.13.491861

Gregory, T. R. (2005). The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* 95, 133–146. doi: 10.1093/aob/mci009

Greilhuber, J., Dolezel, J., Lysák, M. A., and Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms "genome size" and "C-value" to describe nuclear DNA contents. *Ann. Bot.* 95, 255–260. doi: 10.1093/aob/mci019

Greilhuber, J., and Leitch, I. J. (2013). ""Genome size and the phenotype,"," in *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Eds. J. Greilhuber, J. Dolezel and J. F. Wendel (Springer, Vienna), 323–344. doi: 10.1007/978-3-7091-1160-4_20

Grover, C. E., Hawkins, J. S., and Wendel, J. F. (2008). Phylogenetic insights into the pace and pattern of plant genome size evolution. *Genome Dyn.* 4, 57–68. doi: 10.1159/000126006

Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A., and Wendel, J. F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.* 16, 1252–1261. doi: 10.1101/gr.5282906

Heslop-Harrison, J. S., Schwarzacher, T., and Liu, Q. (2023). Polyploidy: its consequences and enabling role in plant diversification and evolution. *Ann. Bot.* 131, 1–10. doi: 10.1093/aob/mcac132

Hozza, M., Vinař, T., and Brejová, B. (2015). "How big is that genome? Estimating genome size and coverage from k-mer abundance spectra," in String Processing and Information Retrieval *Lecture Notes in Computer Science*. Eds. C. Iliopoulos, S. Puglisi and E. Yilmaz (Springer International Publishing, Cham), 199–209. doi: 10.1007/978-3-319-23826-5_20

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13, R3. doi: 10.1186/gb-2012-13-1-r3

Jiao, Y., Li, J., Tang, H., and Paterson, A. H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots[W]. *Plant Cell* 26, 2792–2802. doi: 10.1105/tpc.114.127597

Kelly, L., Leitch, A., Fay, M., Renny-Byfield, S., Pellicer, J., Macas, J., et al. (2012). Why size really matters when sequencing plant genomes. *Plant Ecol. Divers. - Plant Ecol. Divers.* 5. doi: 10.1080/17550874.2012.716868

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., et al. (2015). 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. doi: 10.1111/nph.13471

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2, 231–239. doi: 10.1016/0888-7543(88)90007-9

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945. doi: 10.1093/genetics/166.2.935

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. doi: 10.48550/arXiv.1303.3997

Li, F., and Harkess, A. (2018). A guide to sequence your favorite plant genomes. *Appl. Plant Sci.* 6, e103. doi: 10.1002/aps3.1030

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2020). Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. doi: 10.48550/arXiv.1308.2012

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., et al. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* 45, 884–890. doi: 10.1038/ng.2678

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.. *EMBnet. journal.* 17 (1), 10–12. doi: 10.14806/ej.17.1.200

Mirsky, A. E., and Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* 34, 451–462. doi: 10.1085/jgp.34.4.451

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121. doi: 10.1093/nar/gkt263

Morse, A. M., Peterson, D. G., Islam-Faridi, M. N., Smith, K. E., Magbanua, Z., Garcia, S. A., et al. (2009). Evolution of genome size and complexity in pinus. *PLoS One* 4, e4332. doi: 10.1371/journal.pone.0004332

Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* 49, 490–496. doi: 10.1038/ng.3813

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., et al. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods* 16, 1. doi: 10.1186/s13007-019-0534-5

Noel, B., Denoeud, F., Rouan, A., Buitrago-López, C., Capasso, L., Poulain, J., et al. (2023). Pervasive tandem duplications and convergent evolution shape coral genomes. *Genome Biol.* 24, 123. doi: 10.1186/s13059-023-02960-7

Ohri, D. (2021). Variation and evolution of genome size in gymnosperms. *Silvae Genet.* 70, 156–169. doi: 10.2478/sg-2021-0013

Olofsson, J. K., Cantera, I., Van de Paer, C., Hong-Wa, C., Zedane, L., Dunning, L. T., et al. (2019). Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Mol. Ecol. Resour.* 19, 877–892. doi: 10.1111/1755-0998.13016

Pellicer, J., Fernández, P., Fay, M. F., Michálková, E., and Leitch, I. J. (2021). Genome size doubling arises from the differential repetitive DNA dynamics in the genus heloniopsis (Melanthiaceae). *Front. Genet.* 12. doi: 10.3389/fgene.2021.726211

Pellicer, J., Hidalgo, O., Dodsworth, S., and Leitch, I. J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes* 9, 88. doi: 10.3390/genes9020088

Pellicer, J., and Leitch, I. J. (2014). The application of flow cytometry for estimating genome size and ploidy level in plants. *Methods Mol. Biol. Clifton NJ* 1115, 279–307. doi: 10.1007/978-1-62703-767-9_14

Pellicer, J., and Leitch, I. J. (2020). 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. doi: 10.1111/nph.16261

Pfenninger, M., Schönnenbeck, P., and Schell, T. (2022). ModEst: Accurate estimation of genome size from next generation sequencing data. *Mol. Ecol. Resour.* 22, 1454–1464. doi: 10.1111/1755-0998.13570

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., and Maddison, D. R. (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). *G3 GenesGenomesGenetics* 10, 3047–3060. doi: 10.1534/g3.120.401028

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Pouchon, C., Boyer, F., Roquet, C., Denoeud, F., Chave, J., Coissac, E., et al. (2022). ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications. *Mol. Ecol. Resour.* 22, 2018–2037. doi: 10.1111/1755-0998.13584

Pucker, B. (2019). Mapping-based genome size estimation 607390. doi: 10.1101/607390

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi: 10.1038/s41467-020-14998-3

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., et al. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* 11, 414–428. doi: 10.1016/j.molp.2018.01.002

Sanmiguel, P., and Bennetzen, J. L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* 82, 37–44. doi: 10.1006/anbo.1998.0746

Sarmashghi, S., Balaban, M., Rachtman, E., Touri, B., Mirarab, S., and Bafna, V. (2021). Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT. *PLoS Comput. Biol.* 17, e100944. doi: 10.1371/journal.pcbi.1009449

Schmuths, H., Meister, A., Horres, R., and Bachmann, K. (2004). Genome size variation among accessions of Arabidopsis thaliana. *Ann. Bot.* 93, 317–321. doi: 10.1093/aob/mch037

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Sliwinska, E., Loureiro, J., Leitch, I. J., Šmarda, P., Bainard, J., Bureš, P., et al. (2022). Application-based guidelines for best practices in plant flow cytometry. *Cytometry A* 101, 749–781. doi: 10.1002/cyto.a.24499

Smyčka, J., Roquet, C., Boleda, M., Alberti, A., Boyer, F., Douzet, R., et al. (2022). Tempo and drivers of plant diversification in the European mountain system. *Nat. Commun.* 13, 2750. doi: 10.1038/s41467-022-30394-5

Sun, H., Ding, J., Piednoël, M., and Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* 34, 550–557. doi: 10.1093/bioinformatics/btx637

Sun, M., Pang, E., Bai, W.-N., Zhang, D.-Y., and Lin, K. (2023). ploidyfrost: Reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs. *Mol. Ecol. Resour.* 23, 499–510. doi: 10.1111/1755-0998.13720

Temsch, E. M., Koutecký, P., Urfus, T., Šmarda, P., and Doležel, J. (2022). Reference standards for flow cytometric estimation of absolute nuclear DNA content in plants. *Cytometry A* 101, 710–724. doi: 10.1002/cyto.a.24495

Vestek, A., Slovák, M., Weiss-Schneeweiss, H., Temsch, E. M., Luković, J., Kučera, J., et al. (2019). Morpho-anatomical differentiation and genome size variation in three ploidy levels within the B7 cytotype of Prospero autumnale (Hyacinthaceae) complex from the Balkan Peninsula and Pannonian Basin. *Plant Syst. Evol.* 305, 597–609. doi: 10.1007/s00606-019-01581-7

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinforma. Oxf. Engl.* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Weiss-Schneeweiss, H., and Schneeweiss, G. M. (2013). ""Karyotype diversity and evolutionary trends in angiosperms,"," in *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Eds. J. Greilhuber, J. Dolezel and J. F. Wendel (Springer, Vienna), 209–230. doi: 10.1007/978-3-7091-1160-4_13

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. doi: 10.1186/s13059-019-1891-0