Mapping-based Genome Size Estimation

Shakunthala Natarajan[1], Jessica Gehrke[1], Boas Pucker[1*]

[1] Plant Biotechnology and Bioinformatics, Institute of Plant Biology & BRICS, TU Braunschweig, Mendelssohnstrasse 4, 38106 Braunschweig, Germany

* corresponding author: b.pucker@tu-braunschweig.de

SN: 0000-0002-4713-2380

JG: 0009-0002-0967-766X

BP: 0000-0002-3321-7471

Abstract

While the size of chromosomes can be measured under a microscope, obtaining the exact size of a genome remains a challenge. Biochemical methods and k-mer distribution-based approaches allow only estimations. An alternative approach to estimate the genome size based on high contiguity assemblies and read mappings is presented here. Analyses of *Arabidopsis thaliana* and *Beta vulgaris* data sets are presented to show the impact of different parameters. *Oryza sativa, Brachypodium distachyon*, *Solanum lycopersicum*, *Vitis vinifera*, and *Zea mays* were also analyzed to demonstrate the broad applicability of this approach. Further, MGSE was also used to analyze *Eschericia coli, Saccharomyces cerevisiae,* and *Caenorhabditis elegans* datasets to show its utility beyond plants. Mapping-based Genome Size Estimation (MGSE) and additional

scripts are available on GitHub: https://github.com/bpucker/MGSE. MGSE predicts genome sizes based on short reads or long reads requiring a minimal coverage of 5-fold.

Introduction

Nearly all parts of a plant are now tractable to measure, but assessing the size of a plant genome is still challenging. Although chromosome sizes can be measured under a microscope [1], the combined length of all DNA molecules in a single cell is usually unknown. About 25 years after the release of the first *Arabidopsis thaliana* genome sequence, this holds even true for one of the most important model species. Initially, biochemical methods like reassociation kinetics [2], Feulgen photometry [3], quantitative gel blot hybridization [4], southern blotting [5], and flow cytometry [6, 7] were applied. Unfortunately, these experimental methods rely on a reference genome [8]. The rise of next generation sequencing technologies [9] enabled new approaches based on k-mer profiles or the counting of unique k-mers [10, 11]. JellyFish [11], Kmergenie [12], Tallymer [13], Kmerlight [14], and genomic character estimator (gce) [15] are dedicated tools to analyze k-mers in reads. But such k-mer-based estimation methods require a high sequencing coverage [16]. Next, genome sizes can be estimated based on unique k-mers or a complete k-mer profile. Many assemblers like SOAPdenovo [17] and ALLPATHS-LG [18] perform an internal estimation of the genome size to infer an expected assembly size. Dedicated tools for genome size estimation like GenomeScope2 [19, 20] and findGSE [21] were developed. Although the authors considered and addressed a plethora of issues with real data [19], results from different sequencing data sets for the same species can vary. While some proportion of this variation can be attributed to accession-specific differences as described e.g. for *A. thaliana* [21, 22], specific properties of a sequencing library might have an impact on the estimated genome size. For example, high levels of bacterial or fungal contamination could bias the result if not removed prior to the estimation process. Due to high accuracy requirements, k-mer-based approaches are

usually restricted to high quality short reads and cannot be applied to long reads of third generation sequencing technologies. The rapid development of long read sequencing technologies enables high contiguity assemblies for almost any species and is therefore becoming the standard for genome sequencing projects [23–25]. Nevertheless, some highly repetitive regions of plant genomes like nucleolus organizing region (NOR) and centromeres remain usually unassembled [22, 26, 27]. Therefore, the genome size cannot be inferred directly from the assembly size, but the assembly size can be considered a lower boundary when estimating genome sizes.

Extreme genome size estimates of *A. thaliana,* for example, 70 Mbp [2] or 211 Mbp [28], have been proven to be inaccurate based on insights from recent assemblies [22, 27, 29–32]. However, various methods still predict haploid genome sizes between 125 Mbp and 165 Mbp for *A. thaliana* accessions [29, 33–35]. Substantial technical variation is observed not only between methods, but also between different labs or instruments [36]. As described above, extreme examples for *A. thaliana* display 3 fold differences with respect to the estimated genome size. Since no assembly represents the complete genome, the true genome size remains unknown. An empirical approach, i.e. running different tools and comparing the results, might be a suitable strategy.

This work presents a conceptually different method for the estimation of genome sizes based on the mapping of reads to a high contiguity assembly. Mapping-based Genome Size Estimation (MGSE) is a Python script which processes the coverage information of a read mapping and predicts the size of the underlying genome. Since MGSE is a mapping-based method, it requires a genome sequence as reference for the read mapping process. However, this is not a limitation. The reads used for the genome size estimation, could be used for the assembly. We anticipate that future genome sequencing projects will utilize long read sequencing technologies and would easily generate assemblies of high quality which are more than appropriate for MGSE. Since MGSE relies on read mapping, it is able to support genome size estimations based on long read

datasets unlike the existing kmer-based tools. Further, MGSE's applicability to long read datasets, also helps tackle the issue of highly repetitive regions that interfere in genome size estimation. This is because long reads can span or cover the entire length of repetitive regions, ensuring correct mapping of reads and accurate coverage calculation required for an optimal genome size estimation. MGSE is an orthogonal approach to the existing tools for genome size estimation with different challenges and advantages. It is also suitable for both short and long reads obtained from different sequencing technologies, making it a broadly applicable tool in plant genomics and beyond.

Methods

Data sets

Sequencing data sets of the *A. thaliana* accessions Columbia-0 (Col-0) [32, 37–42] and Niederzenz-1 (Nd-1) [35] as well as several *Beta vulgaris* accessions [43–45] were retrieved from the Sequence Read Archive (AdditionalFile1). Only the paired-end fraction of the two included Nd-1 mate pair libraries was included in this analysis. Genome assembly versions TAIR9 [46], AthNd-1_v2 [27], and RefBeet v1.5 [43, 47] served as references in the read mapping process. The *A. thaliana* assemblies, TAIR9 and Ath-Nd-1_v2, already included plastome and chondrome sequences. Plastome (KR230391.1, [48]) and chondrome (BA000009.3, [49]) sequences were added to RefBeet v1.5 to allow proper placement of respective reads.

Genome sequences of *Brachypodium distachyon* strain Bd21 (GCF_000005505.3 [50]), *Solanum lycopersicum* (GCA_002954035.1 [51]), *Vitis vinifera* cultivar Chardonnay (QGNW01000001.1 [52]), *Oryza sativa* ssp. *japonica* cultivar Nipponbare (GCA_034140825.1 [53]), *Zea mays* cultivar

DK105 (GCA_003709335.1 [54]), *Fragaria x ananassa* cultivar benihoppe (GCA_034370585.1 [55]), *Gossypium hirsutum* (GCF_007990345.1 [56]), *Saccharomyces cerevisiae* strain S288C (GCF_000146045.2, [57]), *Eschericia coli* strain K-12 (GCF_000005845.2, [58]), and *Caenorhabditis elegans* strain Bristol N2 (GCF_000002985.6, [59]) were retrieved from the NCBI. Corresponding read data sets of *Brachypodium distachyon* ([50]), *Solanum lycopersicum* [51, 60–62], *Vitis vinifera* [52], *Oryza sativa* [53, 63], *Zea mays* ([54]), *Fragaria x ananassa*, *Gossypium hirsutum*, *Saccharomyces cerevisiae* [64], *Eschericia coli* [65], and *Caenorhabditis elegans* [66] were retrieved from the Sequence Read Archive (AdditionalFile1). These read datasets were chosen based on similarly analyzed material to the respective reference genome sequence. This was done to ensure that the genome size estimation is representative for the reference strain and is not deviating due to structural differences between different strains.

Genome size estimation

JellyFish2 v2.2.4 [11] was applied for the generation of k-mer profiles which were subjected to GenomeScope2 [19]. Selected k-mer sizes ranged from 19 to 25. Results of different sequencing data sets and different k-mer sizes per accession were compared. Genomic character estimator (gce) [15] and findGSE [21] were applied to infer genome sizes from the k-mer histograms. If tools failed to predict a value or if the prediction was extremely unlikely, values were masked to allow meaningful comparison and accommodation in one figure. The number of displayed data points (i.e. successful predictions) is consequently a quality indicator.

Mapping-based genome size estimation

Despite some known biases [67–69], the underlying assumption of MGSE is a nearly random fragmentation of the DNA and thus an equal distribution of sequencing reads over the complete sequence. If the sequencing coverage per position (C) is known, the genome size (N) can be

calculated by dividing the total amount of sequenced bases (L) by the average coverage value: N = L / C. The working concept of MGSE is explained in Fig. 1.

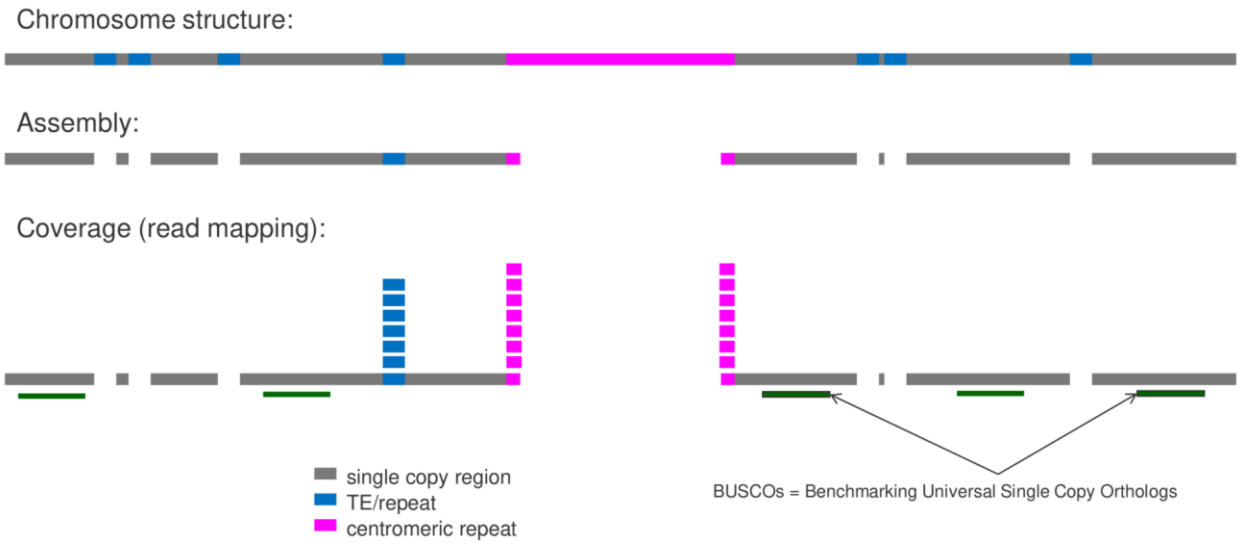


Figure 1: Concept diagram showing the working logic behind read mapping-based genome size estimation. An assembly is required as a basis for the analysis, but does not have to be perfect.

Underrepresented repeats and other regions like the unresolved paralogs of the SEC10 gene in the *Arabidopsis thaliana* genome sequence [70], display a higher coverage, because reads originating from different genomic positions are mapped to the same sequence. The accurate identification of the average coverage is crucial for a precise genome size calculation. Chloroplastic and mitochondrial sequences account for a substantial proportion of reads in sequencing data sets, while contributing very little size compared to the nucleome. Therefore, sequences with very high coverage value, i.e., plastome and chondrome sequences are included during the mapping phase to allow correct placement of reads, but can be excluded from MGSE in a later stage. A user-provided list of reference regions is used to calculate the median or mean coverage based on all positions in these specified regions. Benchmarking Universal Single Copy Orthologs (BUSCO) [71, 72] can be deployed to identify such a set of *bona fide* single copy genes which should serve as suitable regions for the average coverage calculation. Since BUSCO is

frequently applied to assess the completeness of a genome assembly, these files might be already available to users. GFF3 files generated by BUSCO can be concatenated and subjected to MGSE. As some BUSCOs might occur with more than one copy, MGSE provides an option to reduce the predicted gene set to the actual single copy genes among all identified BUSCOs.

For short reads, BWA MEM v0.7 [73] was applied for the read mapping and MarkDuplicates (Picard tools v2.14) [74] was used to filter out reads originating from PCR duplicates. For long reads, minimap2 v2.24 [75, 76] was used for read mapping. Next, a previously described Python script [77] was deployed to generate coverage files from the BAM files, which provides information about the number of aligned sequencing reads covering each position of the reference sequence. This process involves bedtools v2.30.0 [78]. Finally, MGSE v3.1 (https://github.com/bpucker/MGSE) was run on these coverage files to predict genome sizes independently for each data set.

Coverage threshold analysis

MGSE relies on coverage calculation for estimating the genome sizes. Hence, it is important to know the minimum coverage that datasets given to MGSE must have for getting optimal results. For this, all the long and short read datasets of *Arabidopsis thaliana*, described earlier, were taken and sub-sampled into FASTQ files containing varying percentages of the reads - 100%, 75%, 50%, 25%, 10%, 7.5%, 5%, 2.5%, 1%, and 0.5% using seqtk [79]. Then each of these different read files across all the samples were given to MGSE to obtain genome size estimates. The number of sequenced bases in the read files given to MGSE and the genome size estimates of MGSE were correlated to determine the minimum coverage of long and short read datasets that MGSE can process. The above sub-sampled files were also given to GenomeScope2 to obtain a

comparative coverage threshold value for short read datasets that can be handled by this kmer-based tool.

Runtime analysis

Runtime is an important factor when users determine the resources needed for using a particular tool. Here, runtime analyses were conducted by correlating the number of sequenced bases in the read files given to MGSE and the total time taken for a complete run including indexing, read mapping, coverage calculation, and the final genome size prediction. Since minimap2 is slightly more efficient than BWA-MEM, some long read datasets were subjected to MGSE runs with a single CPU. Thereafter, the number of CPUs for the MGSE runs on short and long reads was set to be 10, to evaluate the runtimes with moderate resources. Since read mapping is already a part of most genome assembly processes, users might already have BAM files at hand that could serve as input for MGSE. Therefore, another set of runtime analyses were conducted by correlating the size of the input datasets and only the times taken for coverage calculation and genome size prediction. These MGSE runtimes were then compared with GenomeScope2 runtimes for the same datasets carried out with 10 CPUs to benchmark and understand the performance of MGSE in direct comparison to a kmer-based tool.

Results & discussion

Arabidopsis thaliana genome size

MGSE was deployed to calculate the genome size of the two *A. thaliana* accessions Col-0 and Nd-1 (Fig. 2). In order to identify the best reference region set for the average coverage calculation, different reference region sets were tested. Manually selected single copy genes, all

protein encoding genes, all protein encoding genes without transposable element related genes, only exons of these gene groups, and BUSCOs were evaluated (AdditionalFile2). The results were compared against predictions from GenomeScope2, gce, and findGSE for k-mer sizes 19, 21, 23, and 25.

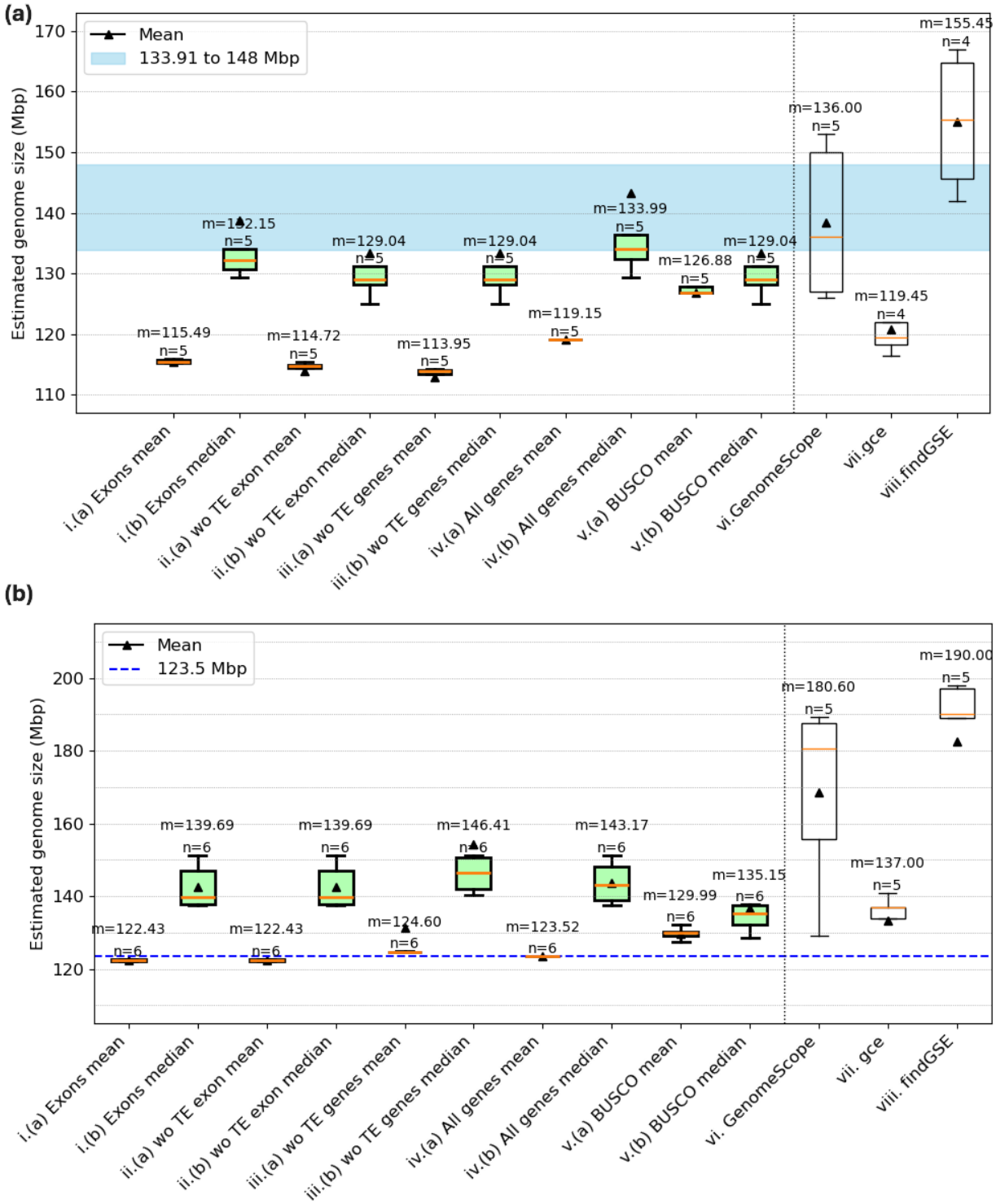


Fig. 2: Comparison of *Arabidopsis thaliana* genome size estimations. Genome sizes of the *A. thaliana* accessions Col-0 (a) and Nd-1 (b) were predicted by MGSE, GenomeScope2, gce, and findGSE. Different MGSE approaches were evaluated, differing by the set of regions for the

average coverage calculation (e.g. all genes) and the methods for the calculation of this value (mean/median). Multiple read data sets (n) were analyzed by each tool/approach to infer an average genome size given as median (m, yellow line) and mean (black triangles), transposable elements = TE, without = wo. The blue region in (a) shows the expected genome size range. It has the near complete assembly size of Col-0 [80] as the lower boundary and one of the largest reported assembly sizes of *Arabidopsis thaliana* [81] as the upper boundary (AdditionalFile8); The blue line in (b) shows a high quality and the largest reported assembly size for Nd-1 till date.

Many estimation approaches predict a genome size for Col-0 that is below the largest reported assembly size of 148 Mbp for *Arabidopsis thaliana* [81] and display substantial variation between samples. The BUSCO-based approaches with MGSE appeared promising due to low variation between different samples and a prediction close to the average sizes of almost complete Col-0 genome sequences. GenomeScope2 predicted a slightly higher genome size with greater variation between samples, while gce reported consistently much smaller values. findGSE predicted on average a substantially larger genome size. Finally, sample sizes below five (for Col-0) and below six (for Nd-1) indicated that prediction processes failed for individual approaches e.g. due to insufficient read numbers.

Due to low variation between different samples and a plausible average genome size, the BUSCO-based approaches appeared promising for Nd-1 as well (Fig. 2b). The reported assembly size of 123.5 Mbp is known to be an underestimation of the true genome size of the Nd-1 accession [27]. The genome size estimation of about 135 Mbp inferred for Nd-1 through the BUSCO approach of MGSE is also slightly below previous estimations of about 146 Mbp [35]. Approximately 123.5 Mbp are assembled into pseudochromosomes which do not contain complete NORs or centromeric regions [27]. Based on the read coverage of the assembled 45S rDNA units, the NORs of Nd-1 are expected to account for approximately 2-4 Mbp [35].

Centromeric repeats which are only partially represented in the genome assembly [27] account for up to 11 Mbp [35]. In summary, the Nd-1 genome size is expected to be around 138-140 Mbp. The single-copy BUSCO genes emerged as the best set of reference regions for MGSE based on further analyses using the Ath-Nd1_v2 assembly. In summary, MGSE (considering the BUSCO median estimation) and gce predicted reasonable genome sizes for Nd-1. The average predictions by GenomeScope2 and findGSE are very unlikely, as they contradict most estimations of *A. thaliana* genome sizes [6, 21, 27, 35].

MGSE was also used to estimate the genome size of the Col-0 accession using long read data sets of 14 GABI-Kat lines [32] (Fig. 3, AdditionalFile3). Since estimates relying on single copy BUSCO genes as reference regions were identified as the best approach in previous *A. thaliana* analyses, the same strategy was applied to datasets of all these GABI-Kat lines.

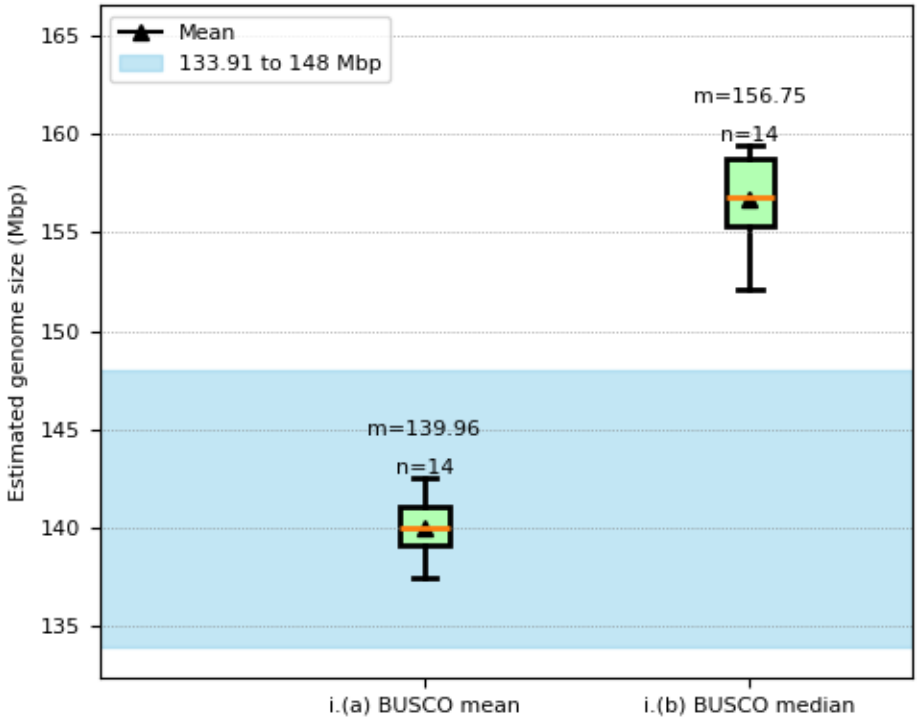


Fig. 3: MGSE estimations using long read sequencing data sets of 14 GABI-Kat lines (Col-0 accession). The blue region shows the expected genome size range with the near complete assembly size of Col-0 [80] as the lower boundary and one of the largest reported assembly sizes of *Arabidopsis thaliana* [81] as the upper boundary (AdditionalFile8).

The MGSE BUSCO mean fits well into the expected genome size range of Col-0, while the median estimate is slightly higher than the largest reported assembly size considered here. This deviation could be attributed to large numbers of InDels in long reads. There are a number of methods to reduce this bias in the input long read data, like the use of PacBio HiFi technology [82] and application of HERRO-correction to ONT reads [83], which could help improve the MGSE estimates. Nevertheless, the estimates still fall into the acceptable limits for *Arabidopsis* genome size estimates. This depicts that MGSE is able to process long read datasets unlike the kmer-based approaches, making it valuable as long read sequencing is gaining importance.

The feasibility of MGSE was further demonstrated by estimating the genome sizes of 1,028 *A. thaliana* accessions (Fig. 4, AdditionalFile4) which were analyzed by re-sequencing as part of the 1001 genome project [84]. Most predictions by MGSE v0.55 are between 120 Mbp and 160 Mbp, while all other tools predict most genome sizes between 120 Mbp and 200 Mbp with some outliers showing very small or very large genome sizes. MGSE differs from all three tools when it comes to the number of failed or extremely low genome size predictions. All k-mer-based approaches predicted genome sizes below 50 Mbp, which are most likely artifacts, possibly due to very low sequencing coverage. This comparison revealed systematic differences between findGSE, gce, and GenomeScope2 with respect to the average predicted genome size. findGSE tends to predict larger genome sizes than gce and GenomeScope2. Very large genome sizes could have biological explanations like polyploidization events.

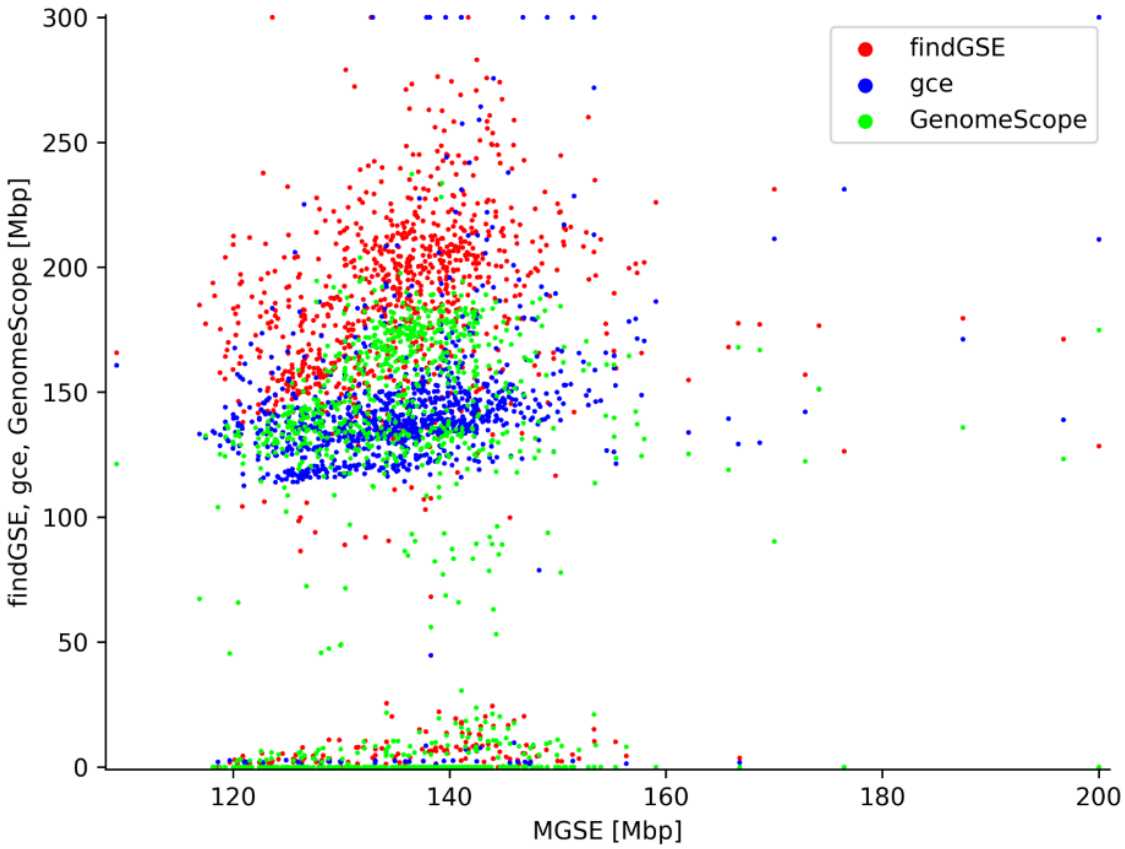


Fig. 4: Genome size estimations of *Arabidopsis thaliana* accessions. MGSE, findGSE, gce, and GenomeScope2 were deployed to predict the genome sizes of 1,028 *A. thaliana* accessions based on sequence read data sets (AdditionalFile4). Extreme outliers above 200 Mbp (MGSE) or 300 Mbp (other tools) are displayed at the plot edge to allow accommodation of all data points with sufficient resolution in the center.

Beta vulgaris genome size

Different sequencing data sets of *Beta vulgaris* were analyzed via MGSE, GenomeScope2, gce, and findGSE to assess the applicability to larger and more complex genomes (Fig. 5, AdditionalFile5). Different cultivars served as material source for the generation of the analyzed

read data sets. Therefore, minor differences in the true genome size are expected. Moreover, sequence differences like single nucleotide variants, small insertions and deletions, as well as larger rearrangements could influence the outcome of this analysis. The reference sequence RefBeet v1.5 already represented 567 Mbp [43, 47], but an investigation of the repeat content indicates a larger genome size due to a high number of repeats that are not represented in the assembly [85]. The largest reported assembly of the reference accession KWS2320 has a size of 573 Mbp [86] thus all estimations below 573 Mbp can be considered as underestimations. Given that the assembly size of the accession U2Bv is even slightly larger with 596 Mbp [86]. When assuming centromere sizes of only 2-3 Mbp per chromosome, the assembly size of 600 Mbp could be close to the true genome size. Previous genome size estimations reported 714-758 Mbp [6] and 731 Mbp [43]. In summary, it appears likely that the true genome size of sugar beet is between about 600 Mbp and 758 Mbp.

In this study, a few samples were considered to be outliers as these samples showed low coverage values. It is possible that these samples belong to a different subspecies which could explain a low mapping rate against the sugar beet reference. Nevertheless, MGSE, and findGSE performed best in estimating the genome sizes of *B. vulgaris*. findGSE estimated extremely variable values, but mostly around the previously estimated genome sizes [6, 43]. It is noticeable that the findGSE analysis failed for a number of samples. It is also important to note that GenomeScope2 and gce underestimate the genome size. The prediction of less than 600 Mbp is an underestimation, because this value is below the size of available genome sequences. Looking at different MGSE modes, the non-BUSCO approaches gave extremely high estimates and showed high variability between samples, the BUSCO-based approaches were in a plausible range. Therefore, the mean and median-based approaches relying on all genes or just the BUSCO genes as reference regions for the sequencing coverage estimation outperformed all other approaches. When comparing the *A. thaliana* and *B. vulgaris* analyses, the calculation of

an average coverage in all single-copy BUSCO genes appears to be the most promising approach. This aligns well with the assumption that single copy BUSCO genes should appear with exactly one copy in the genome and would be prime regions to infer the average sequencing coverage depth.

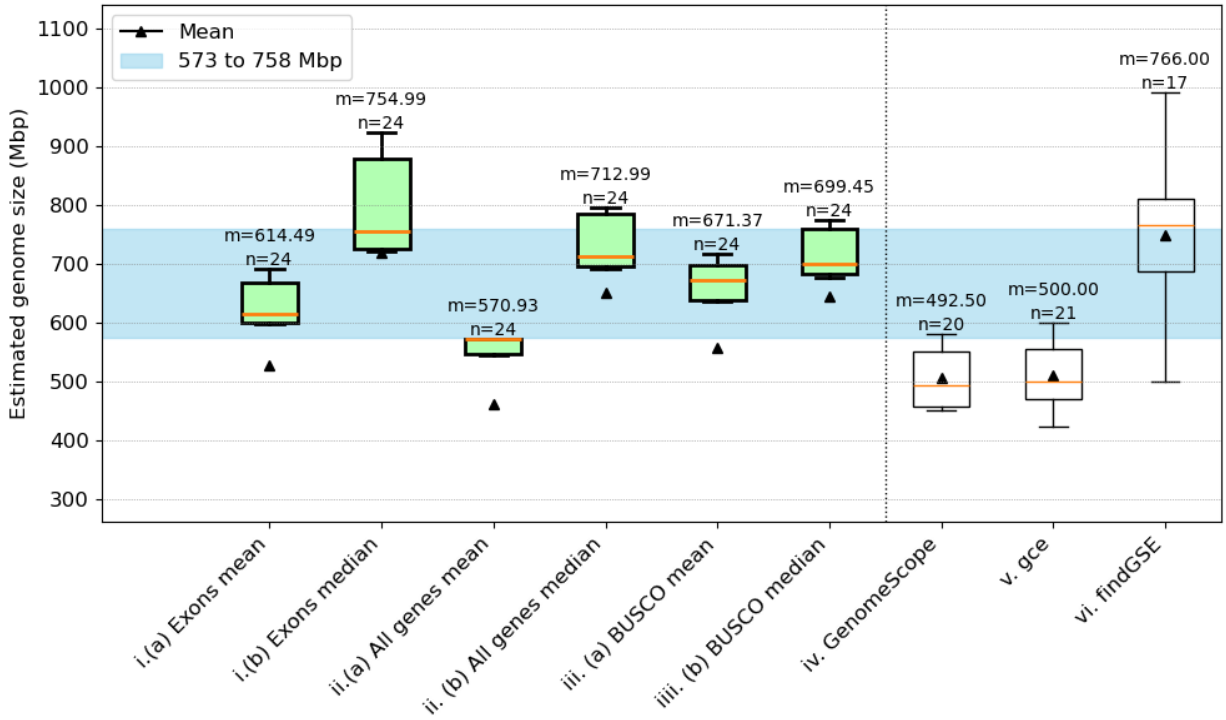


Fig. 5: Comparison of *Beta vulgaris* genome size estimations. The genome size of *B. vulgaris* was predicted by MGSE, GenomeScope2, gce, and findGSE. Different MGSE approaches were evaluated differing by the set of regions for the average coverage calculation (e.g. all genes) and the methods for the calculation of this value (mean/median). Multiple read data sets (n) were analyzed by each tool and approach to infer an average genome size given as median (m, yellow line) and mean (black triangles). The blue region shows the expected genome size range with the largest reported assembly for *Beta vulgaris* [86] as the lower boundary and the largest previously estimated genome size of *Beta vulgaris* [6] as the upper boundary (AdditionalFile8).

Oryza sativa genome size

Rice (*Oryza sativa*) is a major food grain and belongs to the monocot lineage. Given its importance as a food crop, there have been a lot of efforts dedicated to obtaining a complete genome sequence of rice [87]. Recently, a complete genome sequence of *Oryza sativa* ssp. japonica cv. Nipponbare was reported with 385.7 Mbp [53]. It was obtained using a hybrid assembly approach combining reads from Illumina, ONT, and PacBio sequencing technologies with a Hi-C dataset [53].

MGSE was used to estimate the size of the rice genome using reads from different sequencing technologies (Illumina, ONT, and PacBio). Other tools like findGSE, gce, and GenomeScope2 were also used to estimate the genome size using the Illumina reads. As these tools are relying on k-mers, they have only been recommended for the use with short reads. The estimation results using rice Illumina reads are shown in Fig. 6a. The comparative estimations by MGSE using short and long reads are shown in Fig. 6b.

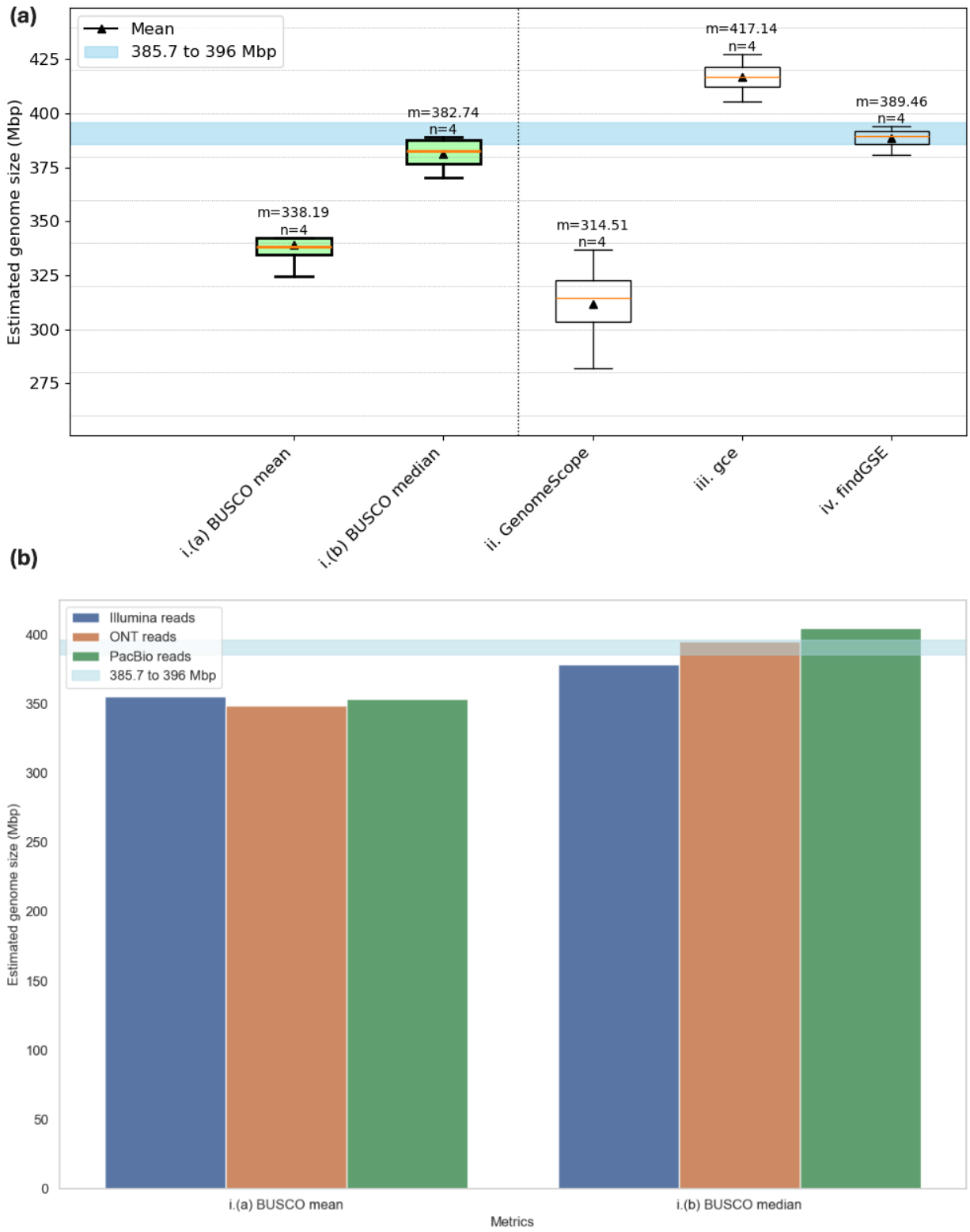


Fig. 6: (a) Comparison of *Oryza sativa* genome size estimations using short reads. (b) Comparison of MGSE genome size estimations on short and long reads of *O. sativa.* The blue

region shows the expected genome size values (AdditionalFile8). It has the size of the near complete genome sequence of *O. sativa* as the lower boundary and the size of the largest reported assembly of *O. sativa* as the upper boundary [88].

While MGSE BUSCO median and findGSE estimate sizes very close to the size of the presumable complete genome sequence, BUSCO mean estimates of MGSE and GenomeScope2 underestimate the genome size. gce slightly overestimates the size.

In contrast to the other tools, MGSE can perform genome size predictions based on long reads. In fact, BUSCO median estimates with MGSE are very close to the expected genome size and they appear to be even more plausible than the estimations based on short reads.

Application to broad taxonomic range of species

After optimization of MGSE on *A. thaliana* (Rosids) and *B. vulgaris* (Caryophyllales), the tool was deployed to analyze data sets of different taxonomic groups thus demonstrating broad applicability. *Brachypodium distachyon* was selected as representative of grasses, *Solanum lycopersicum* represents the Asterids, *Zea mays* was included as monocot species with high transposable element content in the genome, *Vitis vinifera* was selected due to a very high heterozygosity, *Fragaria x ananassa, Gossypium hirsutum* were added to represent polyploid species, *Eschericia coli, Saccharomyces cerevisiae, and Caenorhabditis elegans* were chosen to represent bacterial, fungal and animal genomes. The predictions of MGSE are generally in the same range as the predictions generated by GenomeScope2, gce, and findGSE (AdditionalFile5, AdditionalFile9, AdditionalFile10, AdditionalFile11, and AdditionalFile12). With an average prediction of 290 Mbp as genome size of *B. distachyon*, the MGSE prediction is slightly exceeding the assembly size. findGSE prediction also slightly exceeds the assembly size. GenomeScope2 and gce predict genome sizes below the assembly size (AdditionalFile7). The *Z. mays* genome

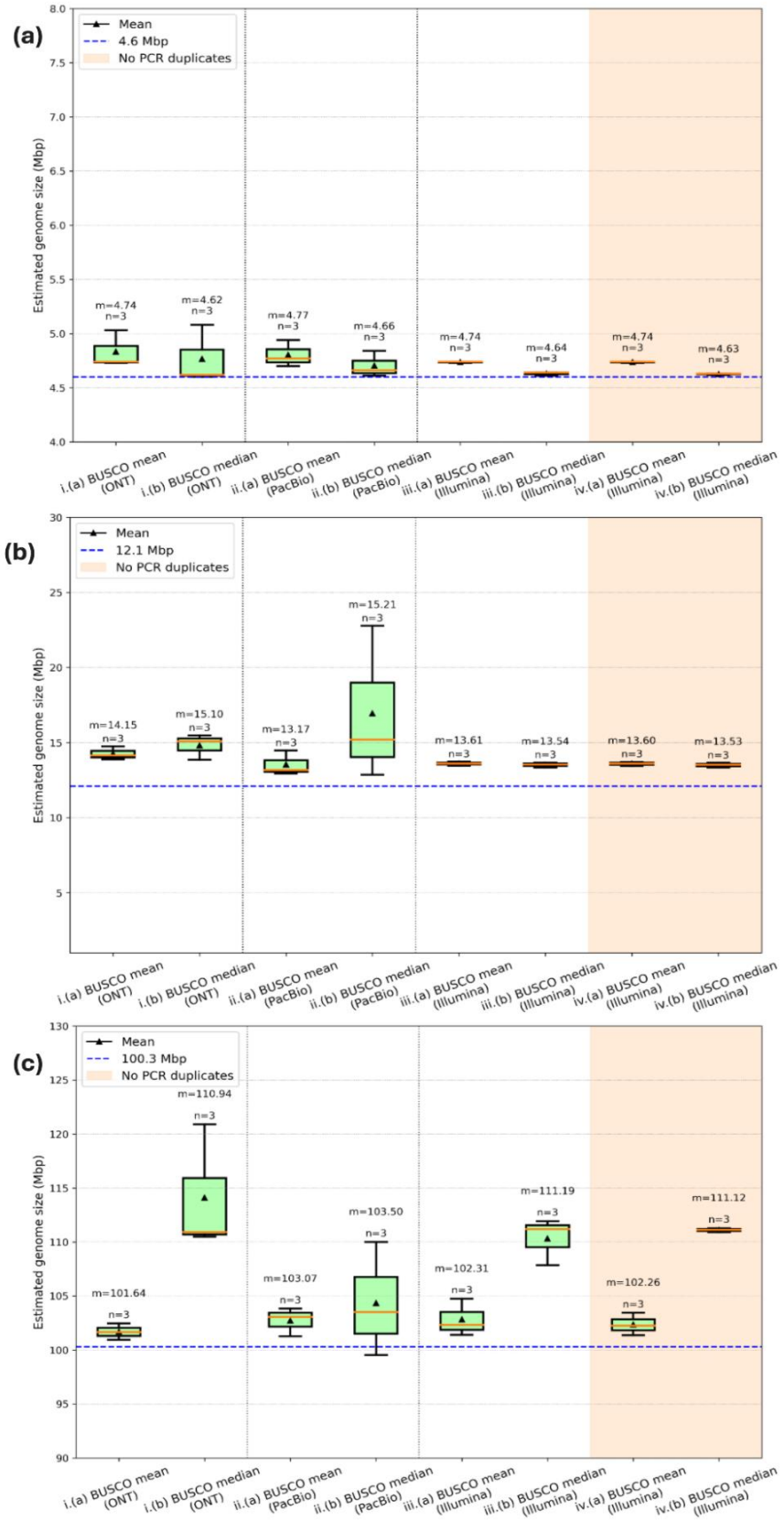
size is underestimated by all four tools. However, MGSE outperforms GenomeScope2 and gce on the analyzed data sets (AdditionalFile8). The *S. lycopersicum* genome size is underestimated by MGSE on most data sets. However, the compared tools failed to predict a genome size for multiple read data sets (AdditionalFile9). MGSE estimates for *V. vinifera* were well within the expected genome size range. findGSE and gce overestimated the genome size. GenomeScope2 underestimated the genome size of *V. vinifera* as evident from the assembly size exceeding the estimation (AdditionalFile10).

To further assess MGSE's applicability to complex polyploid genomes, MGSE was deployed on *Fragaria x ananassa* and *Gossypium hirsutum*. *Fragaria x ananassa* was selected because it is a plant with high ploidy (octoploid) and is also an important crop with global relevance. From the MGSE results, it is evident that, selecting all the genes as well as BUSCO genes, as reference genes for coverage calculation gives optimal results for *Fragaria ananassa*, albeit with a slight overestimation of the median genome size (AdditionalFile13). But it is important to note that polyploid genomes have multiple copies of a region in different subgenomes, resulting in very high coverage values. Hence, it is recommended to select the '--ignore' option to deactivate the blacklisting of high coverage contigs in the case of polyploids. Next, *Gossypium hirsutum*, was selected, as it is an allotetraploid plant with a high degree of repeats and is also an important crop. While MGSE performs optimally when choosing all the genes for coverage calculation and choosing the '--ignore' option for *Gossypium hirsutum*, it fails to predict plausible genome sizes in the other cases (AdditionalFile14). This could be attributed to the high degree of repetitive genomic sequences in upland cotton [89, 90]. Hence, it is suggested to use MGSE with the '--ignore' option, while deploying it on polyploid genomes, especially those with a high number of repeat elements.

Further, it is necessary to note here that MGSE predicts the genome size of the organism, based on the ploidy of the genome assembly provided by the user. To help the user assess the ploidy

level of the assembly that they are providing and to facilitate the selection of optimal parameters for running MGSE, we developed an additional script. 'busco2ploidy.py' that estimates the assembly's ploidy by analyzing BUSCO results (i.e. BUSCO gene duplications) of the assembly. Alternatively, we recommend using tools like smudgeplot [20] which help infer the ploidy before deciding upon the best options for an MGSE run.

Moreover, MGSE is suitable for organisms across a wide taxonomic range as depicted for bacterial (*Eschericia coli*), fungal (*Saccharomyces cerevisiae*) and animal (*Caenorhabditis elegans*) genomes. MGSE estimates for *E. coli* and *C. elegans* aligned well with the expected genome sizes of these organisms (Fig. 7a, 7c). But, MGSE slightly overestimates the yeast genome size when compared to the standard reference assembly size (Fig. 7b). However, there are some studies that reported yeast strains with additional genomic regions compared to the standard reference assembly [91]. Further, some recent studies also report a few genome assembly sizes of yeast strains to exceed the value of 12.1 Mb reported for the standard reference [92], corroborating the slightly higher MGSE estimates for yeast.


(a)
Mean
4.6 Mbp
No PCR duplicates
Estimated genome size (Mbp)
m=4.74 n=3
m=4.62 n=3
m=4.77 n=3
m=4.66 n=3
m=4.74 n=3
m=4.64 n=3
m=4.74 n=3
m=4.63 n=3
i.(a) BUSCO mean (ONT)
i.(b) BUSCO median (ONT)
ii.(a) BUSCO mean (PacBio)
ii.(b) BUSCO median (PacBio)
iii.(a) BUSCO mean (Illumina)
iii.(b) BUSCO median (Illumina)
iv.(a) BUSCO mean (Illumina)
iv.(b) BUSCO median (Illumina)
(b)
Mean
12.1 Mbp
No PCR duplicates
Estimated genome size (Mbp)
m=14.15 n=3
m=15.10 n=3
m=13.17 n=3
m=15.21 n=3
m=13.61 n=3
m=13.54 n=3
m=13.60 n=3
m=13.53 n=3
(c)
Mean
100.3 Mbp
No PCR duplicates
Estimated genome size (Mbp)
m=101.64 n=3
m=110.94 n=3
m=103.07 n=3
m=103.50 n=3
m=102.31 n=3
m=111.19 n=3
m=102.26 n=3
m=111.12 n=3

377

Fig. 7: MGSE genome size estimates of (a) *Eschericia coli* (b) *Saccharomyces cerevisiae* (c) *Caenorhabditis elegans.* The dotted blue lines denote the genome assembly sizes of the standard reference organisms. The y-axis was automatically set to a range that highlights differences between the genome size prediction approaches.

MGSE parameter selection and optimization

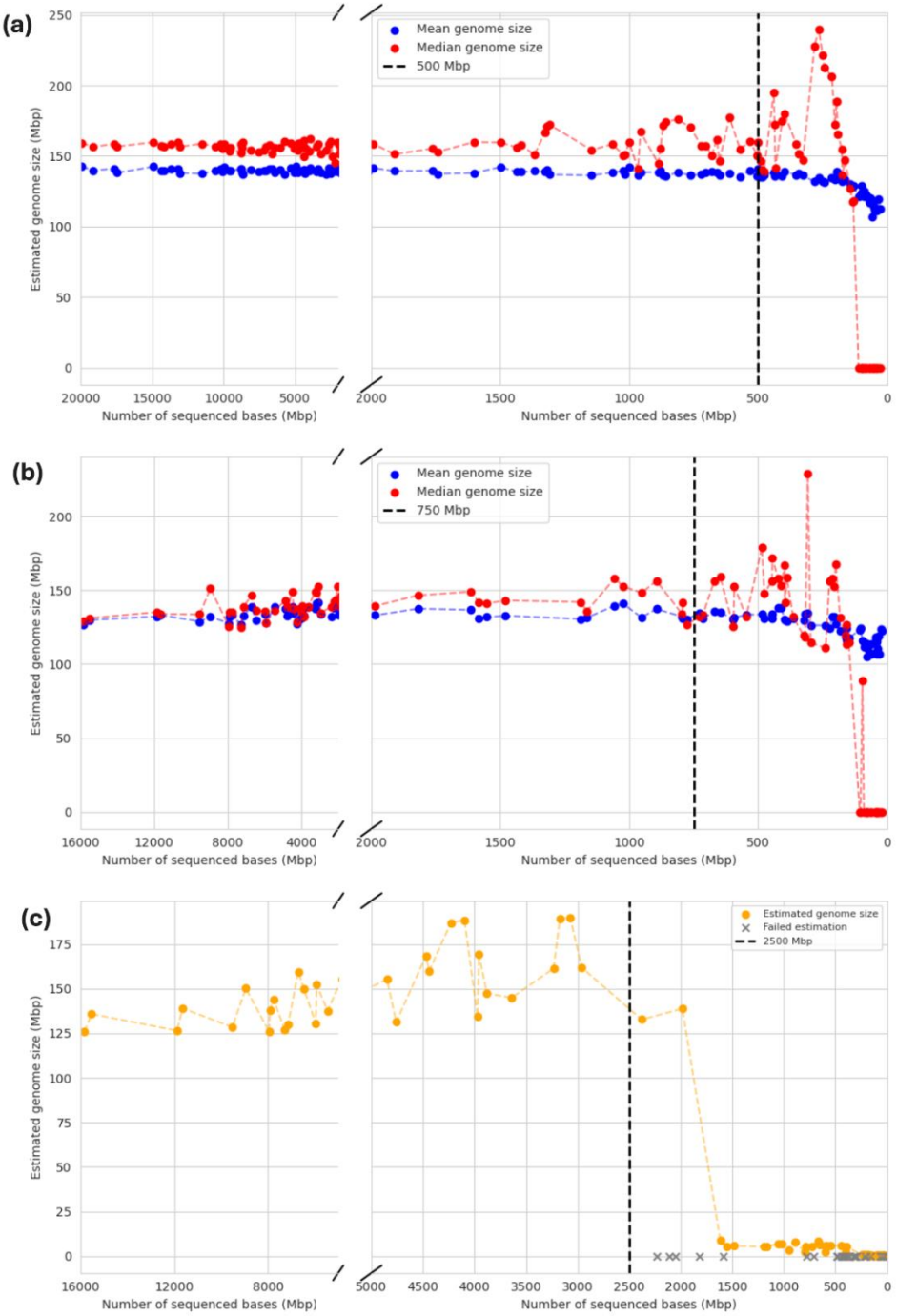
There are some MGSE parameter recommendations for optimal results. First, MGSE can accept different input types including both raw read files along with the reference assembly file, (sorted) BAM files, or coverage files. In many cases, BAM files are already available as a result of a typical genome assembly process which requires a read mapping to check the coverage of contigs as part of the quality control. Available BAM files avoid the computationally intensive step of read mapping as a part of MGSE's execution, making the run more efficient.

Second, selection of the optimal reference regions for coverage calculation is crucial. Based on results obtained for *A. thaliana* and more complex genomes, it became evident that BUSCO genes work well as reference regions in a wide range of application scenarios. Hence, the preference of reference gene sets by the users can be BUSCO genes followed by all the genes. MGSE leverages established methods for coverage calculation by running bedtools thus no optimization is required for this process.

After choosing the regions for coverage calculation, it is necessary to filter out high coverage contigs that could distort the genome size estimations like contaminating sequences from microbes, mitochondrial sequences, and chloroplastic sequences. This blacklisting of high coverage contigs (>1.5x of average coverage) is activated by default, but can be deactivated by

using the '--ignore' flag. However, there are some cases in which it is not recommended to blacklist high coverage contigs, as shown for polyploids and organisms with a high degree of genomic repeats. Including the '--ignore' flag can lead to better results in these cases. The user can run the busco2ploidy.py script to assess the apparent ploidy of the organism before running MGSE, in order to decide upon the optimal parameters. This script provides a frequency distribution plot of the BUSCO genes, and gives the user a 'pseudo ploidy number'. Alternatively, the user can define the '--blacklist_factor' to MGSE instead of using the '--ignore' option. To use this method, the user can set the '--blacklist_factor' to be '1.5 times the pseudo ploidy number', which would heighten the cutoff for high coverage sequences, making it suitable for polyploid genomes and ensuring removal of just those sequences that exhibit unusually high coverage.

Further, coverage of the input datasets to MGSE significantly impacts the MGSE estimates. This is evident from the results of the coverage threshold analysis (AdditionalFile6). For *Arabidopsis*, the minimum number of bases (Mbp) required for optimal MGSE estimates from long reads was found to be 500 Mbp and the minimum number of bases required for optimal MGSE estimates from short reads was found to be 750 Mbp (Fig. 8a, 8b). While the minimum number of bases required for successful prediction by GenomeScope2 on short reads turned out to be 2500 Mbp (Fig. 8c). Given that the expected genome size of A. thaliana is ~150 Mbp, the minimum coverage of datasets for a successful MGSE run based on long reads was found to be ~3X, while that based on short reads was found to be ~5X. These minimum coverage values were significantly lower than the minimum coverage of short read datasets needed for an optimal GenomeScope2 prediction, which was ~17X. MGSE's applicability to both long and short reads coupled with its ability to predict genome sizes based on datasets with low coverage values, makes it a valuable tool in addition to the existing kmer-based tools.


(a)
Mean genome size
Median genome size
500 Mbp
Estimated genome size (Mbp)
Number of sequenced bases (Mbp)
(b)
Mean genome size
Median genome size
750 Mbp
Estimated genome size (Mbp)
Number of sequenced bases (Mbp)
(c)
Estimated genome size
Failed estimation
2500 Mbp
Estimated genome size (Mbp)
Number of sequenced bases (Mbp)

427

Fig. 8: Coverage threshold analyses of (a) MGSE on long read datasets (b) MGSE on short read datasets (c) GenomeScope2 on short read datasets.

Finally, the runtime analyses of MGSE and GenomeScope2 revealed interesting patterns (AdditionalFile7). When comparing MGSE and GenomeScope2 in terms of the total runtime and the size of the input dataset, MGSE had higher runtimes than GenomeScope2 (Fig. 9). However, the computationally intensive step is the read mapping and MGSE could also be started with precomputed BAM files. When only considering the runtimes of coverage calculation and genome size prediction, the MGSE runtimes show a near plateau. For very large datasets, the MGSE and GenomeScope2 runtimes started to converge (Fig. 9). This indicates that MGSE can perform on par or even more efficiently than GenomeScope2 on very large datasets.

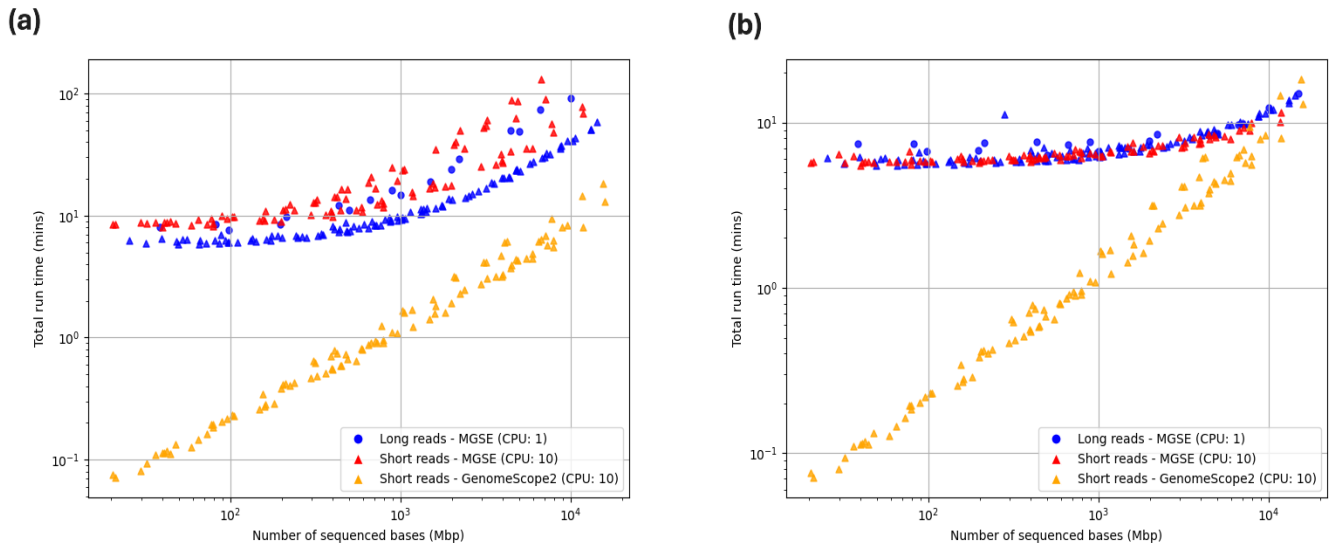


Fig. 9: (a) Comparative analysis between total MGSE and GenomeScope2 runtimes and (b) Comparative analysis between coverage+MGSE execution times and GenomeScope2 total runtimes.

Considerations about performance and outlook

MGSE performs best on a high contiguity assembly and requires a mapping of short or long reads to this assembly. Accurate coverage calculation for each position in the assembly is important and contigs display artificially low coverage values towards the ends. This is caused by a reduction in the number of possible ways in which reads can cover contig ends. The shorter a contig, the more the apparent coverage of this contig is reduced. Given that most assemblies are characterized by N50 values of >10Mbp, this effect can be neglected. Since a read mapping is required as input, MGSE might appear less convenient than classical k-mer-based approaches at first look. However, these input files are already available for many plant species, because such mappings are often part of the assembly (quality control) process [26, 27, 93, 94]. Future genome sequencing projects are likely to generate high continuity assemblies [25] and read mappings in the polishing process.

One advantage of MGSE is the possibility to exclude reads originating from contaminating DNA even if the proportion of such DNA is high. Unless reads from bacterial or fungal contaminations were assembled and included in the reference sequence, the approach can handle such reads without identifying them explicitly. This is achieved by discarding unmapped reads from the genome size estimation. MGSE works best with a high contiguity assembly and assumes all single copy regions of the genome are resolved and all repeats are represented by at least one copy. Although the amount of contamination of or in reads is usually small, such reads are frequently observed due to the high sensitivity of next generation sequencing methods [35, 95–97].

Reads originating from PCR duplicates could impact k-mer profiles and also predictions based on these profiles if not filtered out. After reads are mapped to a reference sequence, read pairs originating from PCR duplicates can be identified and removed based on identical start and end

positions as well as identical sequences. This results in the genome size prediction by MGSE being independent of the library diversity. If the coverage is close to the read length or the length of sequenced fragments, reads originating from PCR duplicates cannot be distinguished from *bona fide* identical DNA fragments. Although MGSE results get more accurate with higher coverage, after exceeding an optimal coverage the removal of apparent PCR duplicates could become an issue. Thus, a substantially higher number of reads originating from PCR-free libraries could be used if duplicate removal is omitted. Depending on the sequencing library diversity, completely skipping the PCR duplicate removal step might be an option for further improvement. As long as these PCR duplicates are mapped equally across the genome, MGSE can tolerate these artifacts. This is in fact shown by the MGSE estimates of *E. coli*, *S. cerevisiae*, and *C. elegans* based on short read datasets. Removal of PCR duplicates from these datasets does not impact the MGSE size estimates (Fig. 7).

All methods are affected by DNA of the plastome and chondrome integrated into the nuclear chromosomes [98, 99]. K-mers originating from these sequences are probably ignored in many k-mer-based approaches, because they appear to originate from the chondrome or plastome, i.e., k-mers occur with very high frequencies. The apparent coverage in the mapping-based calculation is biased due to high numbers of reads which are erroneously mapped to these sequences instead of the plastome or chondrome sequence.

Differences in the GC content of genomic regions were previously reported to have an impact on the sequencing coverage [100, 101]. Both, extremely GC-rich and AT-rich fragments, respectively, are underrepresented in the sequencing output mainly due to biases introduced by PCR [102, 103]. Sophisticated methods were developed to correct coverage values based on the GC content of the underlying sequence [103–105]. The GC content of genes selected as reference regions for the coverage estimation is likely to be above the 36.3% average GC content of plants [77]. This becomes worse when only exons are selected due to the even higher

proportion of coding sequence. Although a species-specific codon usage can lead to some variation, constraints of the genetic code determine a GC content of approximately 50% in coding regions. The selection of a large set of reference regions with a GC content close to the expected overall GC content of a genome would be ideal. However, the overall GC content is unknown as the GC content of regions not represented in the reference sequence is not known and cannot be inferred from the reads. As a result, the average sequencing coverage could be overestimated leading to an underestimation of the genome size. Future investigations would be necessary to develop a correction factor for this GC bias of short reads to further optimize the genome size prediction.

Many plant genomes pose an additional challenge due to recent polyploidy or high heterozygosity. Once high contiguity long read assemblies become available for these complex genomes, a mapping-based approach is feasible. As long as the different haplophases are properly resolved, the assessment of coverage values should reveal a good estimation of the genome size. Even the genomes of species which have recently undergone polyploidization could be investigated with moderate adjustments to the workflow. Reference regions need to be selected to reflect the degree of ploidy in their copy number.

With the widespread adoption of long read sequencing technologies in plant genomics, MGSE can turn out to be an important tool for genome size estimation. The major issue when developing tools for genome size prediction is the absence of a gold standard. In this study, predictions were compared against the best available genome sequences for the respective species. Several of these genome sequences should be very close to a perfect representation of the genome and thus have the potential to reveal the true genome size. Predictions generated by MGSE were constantly close to the sizes of these almost complete genome sequences. Moreover MGSE is able to handle low coverage datasets of both short and long reads, and still return reasonable

genome size estimations. Finally, MGSE is universally applicable to all species and is not restricted to plants.

Data availability

Scripts developed as part of this work and a test data set are freely available on GitHub: https://github.com/bpucker/MGSE (https://doi.org/10.5281/zenodo.13984294). Underlying data sets are publicly available at the NCBI and SRA, respectively.

Supplements

AdditionalFile1: Sequencing data set overview.

AdditionalFile2: *A. thaliana* genome size prediction values for all different approaches.

AdditionalFile3: *A. thaliana* genome size prediction of MGSE based on long reads

AdditionalFile4: *A. thaliana* genome size predictions by MGSE, findGSE, gce, and GenomeScope2.

AdditionalFile5: *B. vulgaris, Zea mays, Oryza sativa, Brachypodium distachyon, Solanum lycopersicum*, *Vitis vinifera, Fragaria x ananassa, Gossypium hirsutum, Eschericia coli, Saccharomyces cerevisiae,* and *Caenorhabditis elegans* genome size prediction values for all different approaches.

AdditionalFile6: Coverage threshold analyses on datasets for MGSE and GenomeScope2

AdditionalFile7: Runtime analyses of MGSE and GenomeScope2

AdditionalFile8: List of near complete/ complete and largest reported genome assembly sizes for the different organisms analyzed in this study

AdditionalFile9: Genome size estimation of *Brachypodium distachyon*.

AdditionalFile10: Genome size estimation of *Zea mays*.

AdditionalFile11: Genome size estimation of *Solanum lycopersicum*.

AdditionalFile12: Genome size estimation of *Vitis vinifera*.

AdditionalFile13: Genome size estimation of *Fragaria x ananassa.*

AdditionalFile14: Genome size estimation of *Gossypium hirsutum*.

Author contributions

BP conceptualized the project, wrote the initial version of MGSE, performed several data analyses, generated figures, and wrote the initial manuscript version. SN updated the MGSE code, conducted additional data analyses, generated figures, and improved the manuscript. JG conducted additional data analyses and validated analyses. All authors read the final version of the manuscript and agreed to its submission.

Conflict of interests

The authors declare that they have no conflict of interests.

Acknowledgements

We are very grateful to the Bioinformatics Resource Facility support team of the CeBiTec for providing computing infrastructure and excellent technical support. Many thanks to members of the Genetics and Genomics of Plants (Bielefeld University) group who contributed to this work by discussion of preliminary results. Many thanks go to Hanna Schilbert, Nathanael Walker-Hale, and Iain Place for helpful comments on the manuscript. We also thank the Plant Biotechnology and Bioinformatics group (TU Braunschweig) for inspiring discussions. SN was supported by the doctoral scholarship of the German Academic Exchange Service (DAAD). This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

References

1. Albini S m. A karyotype of the Arabidopsis thaliana genome derived from synaptonemal complex analysis at prophase I of meiosis. The Plant Journal. 1994;5:665–72.

2. Leutwiler LS, Hough-Evans BR, Meyerowitz EM. The DNA of Arabidopsis thaliana. Molec Gen Genet. 1984;194:15–23.

3. Bennett MD, Smith JB. Nuclear DNA Amounts in Angiosperms. Philosophical Transactions: Biological Sciences. 1991;334:309–45.

4. Francis DM, Hulbert SH, Michelmore RW. Genome size and complexity of the obligate fungal pathogen, Bremia lactucae. Experimental Mycology. 1990;14:299–309.

5. Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I. Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. Proc Natl Acad Sci U S A. 2002;99:14584–9.

6. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 1991;9:208–18.

7. Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot. 2011;107:467–590.

8. Bennett MD, Leitch IJ, Price HJ, Johnston JS. Comparisons with Caenorhabditis (~100 Mb) and Drosophila (~175 Mb) Using Flow Cytometry Show Genome Size in Arabidopsis to be ~157 Mb and thus ~25 % Larger than the Arabidopsis Genome Initiative Estimate of ~125 Mb. Ann Bot. 2003;91:547–57.

9. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010;11:31–46.

10. Li X, Waterman MS. Estimating the Repeat Structure and Length of DNA Sequences Using ℓ-Tuples. Genome Res. 2003;13:1916–22.

11. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.

12. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30:31–7.

13. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9:517.

14. Sivadasan N, Srinivasan R, Goyal K. Kmerlight: fast and accurate k-mer abundance estimation. 2016.

15. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. 2020.

16. Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). G3 Genes|Genomes|Genetics. 2020;10:3047–60.

17. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20:265–72.

18. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences. 2011;108:1513–8.

19. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:2202–4.

20. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432.

21. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics. 2018;34:550–7.

22. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences. 2016;113:E4052–60.

23. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

24. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017;12:213–8.

25. Pucker B, Irisarri I, Vries J de, Xu B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quantitative Plant Biology. 2022;3:e5.

26. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nature Communications. 2018;9:541.

27. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLOS ONE. 2019;14:e0216233.

28. Schmuths H, Meister A, Horres R, Bachmann K. Genome Size Variation among Accessions of Arabidopsis thaliana. Ann Bot. 2004;93:317–21.

29. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.

30. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, et al. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data. 2014;1:140045.

31. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623–30.

32. Pucker B, Kleinbölting N, Weisshaar B. Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. BMC Genomics. 2021;22:599.

33. Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet. 1999;33:479–532.

34. Bevan M, Walsh S. The Arabidopsis genome: a foundation for plant research. Genome Res. 2005;15:1632–42.

35. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11:e0164321.

36. Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, et al. Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. Annals of Botany. 1998;82 suppl_1:17–26.

37. DeFraia CT, Zhang X, Mou Z. Elongator subunit 2 is an accelerator of immune responses in Arabidopsis thaliana. Plant J. 2010;64:511–23.

38. Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability

in Arabidopsis and humans. Genome Res. 2015;25:645–54.

39. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B. The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. Molecular Plant. 2015;8:1651–64.

40. Pellaud S, Bory A, Chabert V, Romanens J, Chaisse-Leal L, Doan AV, et al. WRINKLED1 and ACYL-COA:DIACYLGLYCEROL ACYLTRANSFERASE1 regulate tocochromanol metabolism in Arabidopsis. New Phytol. 2018;217:245–60.

41. Li J, Liang W, Li Y, Qian W. APURINIC/APYRIMIDINIC ENDONUCLEASE2 and ZINC FINGER DNA 3′-PHOSPHOESTERASE Play Overlapping Roles in the Maintenance of Epigenome and Genome Stability. The Plant Cell. 2018;30:1954–70.

42. Wynn E, Purfeerst E, Christensen A. Mitochondrial DNA Repair in an Arabidopsis thaliana Uracil N-Glycosylase Mutant. Plants. 2020;9:261.

43. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature. 2014;505:546–9.

44. Tränkner C, Lemnian IM, Emrani N, Pfeiffer N, Tiwari SP, Kopisch-Obuch FJ, et al. A Detailed Analysis of the BR1 Locus Suggests a New Mechanism for Bolting after Winter in Sugar Beet (Beta vulgaris L.). Frontiers in Plant Science. 2016;7.

45. Funk A, Galewski P, McGrath JM. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. The Plant Journal. 2018;95:659–71.

46. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40 Database issue:D1202–10.

47. Holtgräwe D, Rosleff Sörensen T, Parol-Kryger R, Pucker B, Viehöver P, Kleinbölting N, et al. Low coverage re-sequencing in Beta vulgaris (sugar beet) for anchoring assembly sequences to genomic position. 2017.

48. Stadermann KB, Weisshaar B, Holtgräwe D. SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. BMC Bioinformatics. 2015;16:295.

49. Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNACys(GCA). Nucleic Acids Res. 2000;28:2571–6.

50. Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, et al. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010;463:763–8.

51. Li J, Chitwood J, Menda N, Mueller L, Hutton SF. Linkage between the I-3 gene for resistance to Fusarium wilt race 3 and increased sensitivity to bacterial spot in tomato. Theor Appl Genet. 2018;131:145–55.

52. Roach MJ, Johnson DL, Bohlmann J, Vuuren HJJ van, Jones SJM, Pretorius IS, et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLOS Genetics. 2018;14:e1007807.

53. Shang L, He W, Wang T, Yang Y, Xu Q, Zhao X, et al. A complete assembly of the rice Nipponbare reference genome. Molecular Plant. 2023;16:1232–6.

54. Unterseer S, Seidel MA, Bauer E, Haberer G, Hochholdinger F, Opitz N, et al. European Flint reference sequences complement the maize pan-genome. 2017;:103747.

55. Song Y, Peng Y, Liu L, Li G, Zhao X, Wang X, et al. Phased gap-free genome assembly of octoploid cultivated strawberry illustrates the genetic and epigenetic divergence among subgenomes. Horticulture Research. 2024;11:uhad252.

56. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, et al. Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet. 2020;52:525–33.

57. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 Genes. Science. 1996;274:546–67.

58. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, et al. The Complete Genome Sequence of Escherichia coli K-12. Science. 1997;277:1453–62.

59. Wang X, Zhang L, Zhang L, Wang W, Wei S, Wang J, et al. Effects of excess sugars and lipids on the growth and development of Caenorhabditis elegans. Genes & Nutrition. 2020;15:1.

60. Lang Z, Wang Y, Tang K, Tang D, Datsenka T, Cheng J, et al. Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. Proceedings of the National Academy of Sciences. 2017;114:E4511–9.

61. Corem S, Doron-Faigenboim A, Jouffroy O, Maumus F, Arazi T, Bouché N. Redistribution of CHH Methylation and Small Interfering RNAs across the Genome of Tomato ddm1 Mutants. The Plant Cell. 2018;30:1628–44.

62. Yang Y, Tang K, Datsenka TU, Liu W, Lv S, Lang Z, et al. Critical function of DNA methyltransferase 1 in tomato development and regulation of the DNA methylome and transcriptome. Journal of Integrative Plant Biology. 2019;61:1224–42.

63. Mahesh HB, Shirke MD, Singh S, Rajamani A, Hittalmani S, Wang G-L, et al. Indica rice genome assembly, annotation and mining of blast disease resistance genes. BMC Genomics. 2016;17:242.

64. Zhang X, Liu C-G, Yang S-H, Wang X, Bai F-W, Wang Z. Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. Briefings in Bioinformatics. 2022;23:bbac146.

65. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. PLOS Genetics. 2016;12:e1005854.

66. Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, et al. Recompleting

the Caenorhabditis elegans genome. Genome Res. 2019;29:1009–22.

67. Grokhovsky SL, Il'icheva IA, Nechipurenko DYu, Golovkin MV, Panchenko LA, Polozov RV, et al. Sequence-Specific Ultrasonic Cleavage of DNA. Biophys J. 2011;100:117–25.

68. van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, Toonen P, et al. Systematic biases in DNA copy number originate from isolation procedures. Genome Biology. 2013;14:R33.

69. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. 2014;4:4532.

70. Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, et al. Dissecting a Hidden Gene Duplication: The Arabidopsis thaliana SEC10 Locus. PLOS ONE. 2014;9:e94077.

71. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

72. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 2021;38:4647–54.

73. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013.

74. Broad Institute. Picard Tools. 2019.

75. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

76. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.

77. Pucker B, Brockington SF. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. BMC Genomics. 2018;19:980.

78. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

79. Li H. lh3/seqtk. 2025.

80. Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. A near-complete assembly of an Arabidopsis thaliana genome. Mol Plant. 2022;15:1247–50.

81. Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, et al. A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nat Genet. 2024;56:982–91.

82. Kucuk E, van der Sanden BPGH, O'Gorman L, Kwint M, Derks R, Wenger AM, et al. Comprehensive de novo mutation discovery with HiFi long-read sequencing. Genome Medicine. 2023;15:34.

83. Stanojević D, Lin D, Nurk S, Sessions PF de, Šikić M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. 2024;:2024.05.18.594796.

84. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166:481–91.

85. Kowar T, Zakrzewski F, Macas J, Koblížková A, Viehoever P, Weisshaar B, et al. Repeat Composition of CenH3-chromatin and H3K9me2-marked heterochromatin in Sugar Beet (Beta vulgaris). BMC Plant Biology. 2016;16:120.

86. Sielemann K, Pucker B, Orsini E, Elashry A, Schulte L, Viehöver P, et al. Genomic characterization of a nematode tolerance locus in sugar beet. BMC Genomics. 2023;24:748.

87. Huang C, Chen Z, Liang C. *Oryza* pan-genomics: A new foundation for future rice research and improvement. The Crop Journal. 2021;9:622–32.

88. Song J-M, Xie W-Z, Wang S, Guo Y-X, Koo D-H, Kudrna D, et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. Molecular Plant. 2021;14:1757–67.

89. Shan W, Jiang Y, Han J, Wang K. Comprehensive cytological characterization of the *Gossypium hirsutum* genome based on the development of a set of chromosome cytological markers. The Crop Journal. 2016;4:256–65.

90. Chang X, He X, Li J, Liu Z, Pi R, Luo X, et al. High-quality Gossypium hirsutum and Gossypium barbadense genome assemblies reveal the landscape and evolution of centromeres. Plant Commun. 2023;5:100722.

91. Nijkamp JF, van den Broek M, Datema E, de Kok S, Bosman L, Luttik MA, et al. De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology. Microbial Cell Factories. 2012;11:36.

92. Wang M, Li X, Liu X, Hou X, He Y, Yu J-H, et al. Annotation of 2,507 Saccharomyces cerevisiae genomes. Microbiology Spectrum. 2024;12:e03582-23.

93. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. 2017;27:778–86.

94. Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. Nature Plants. 2018;4:473–84.

95. Kumar S, Blaxter ML. Simultaneous genome sequencing of symbionts and their hosts. Symbiosis. 2011;55:119–26.

96. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC

Biology. 2014;12:87.

97. Strong MJ, Xu G, Morici L, Bon-Durant SS, Baddoo M, Lin Z, et al. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. PLOS Pathogens. 2014;10:e1004437.

98. Ayliffe MA, Scott NS, Timmis JN. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol Biol Evol. 1998;15:738–45.

99. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity (Edinb). 2013;111:314–20.

100. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36:e105.

101. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biology. 2013;14:R51.

102. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology. 2011;12:R18.

103. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

104. Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat Biotechnol. 2016;34:1287–91.

105. Teng M, Irizarry RA. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. Genome Res. 2017;27:1930–8.