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. Mapping-based Genome Size Estimation (MGSE) and additional scripts are available on GitHub: https://github.com/bpucker/MGSE.

Introduction

Nearly all parts of the 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 (Albini, 1994), the combined length of all DNA molecules in a single cell is usually unknown. Almost 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 (Leutwiler *et al.*, 1984), Feulgen photometry (Bennett & Smith, 1991), quantitative gel blot hybridization (Francis *et al.*, 1990), southern blotting (Fransz *et al.*, 2002), and flow cytometry (Arumuganathan & Earle, 1991; Bennett & Leitch, 2011) were applied. Unfortunately, these experimental methods rely on a reference genome (Bennett *et al.*, 2003). The rise of next generation sequencing technologies (Metzker, 2010) enabled new approaches based on k-mer profiles or the counting of unique k-mers (Li & Waterman, 2003; Marçais & Kingsford, 2011). JellyFish (Marçais & Kingsford, 2011), Kmergenie (Chikhi & Medvedev, 2014),

Tallymer (Kurtz *et al.*, 2008), Kmerlight (Sivadasan *et al.*, 2016), and genomic character estimator (gce) (Liu *et al.*, 2020) are dedicated tools to analyze k-mers in reads. But such k-mer-based estimation methods require a relatively higher sequencing coverage (Pflug *et al.*, 2020). Next, genome sizes can be estimated based on unique k-mers or a complete k-mer profile. Many assemblers like SOAPdenovo (Li *et al.*, 2010) and ALLPATHS-LG (Gnerre *et al.*, 2011) perform an internal estimation of the genome size to infer an expected assembly size. Dedicated tools for genome size estimation like GenomeScope (Vurture *et al.*, 2017) and findGSE (Sun *et al.*, 2018) were developed. Although the authors considered and addressed a plethora of issues with real data (Vurture *et al.*, 2017), 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* (Zapata *et al.*, 2016; Sun *et al.*, 2018), 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 (Goodwin *et al.*, 2016; Mardis, 2017; Pucker *et al.*, 2022). Nevertheless, some highly repetitive regions of plant genomes like nucleolus organizing region (NOR) and centromeres remain usually unassembled (Zapata *et al.*, 2016; Michael *et al.*, 2018; Pucker *et al.*, 2019). 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 (Leutwiler *et al.*, 1984) or 211 Mbp (Schmuths *et al.*, 2004) have been proven to be inaccurate based on insights from recent assemblies (The Arabidopsis Genome Initiative, 2000; Kim *et al.*, 2014; Berlin *et al.*, 2015; Zapata *et al.*, 2016; Pucker *et al.*, 2019, 2021). However, various methods still predict genome sizes between 125 Mbp and 165 Mbp for diploid *A. thaliana* accessions (Kumar & Bennetzen, 1999; The Arabidopsis Genome Initiative, 2000; Bevan & Walsh, 2005; Pucker *et al.*, 2016). Substantial technical variation is observed not only between methods, but also between different labs or instruments (Doležel *et al.*, 1998). As described above, extreme examples for *A. thaliana* display 3 fold differences with respect to the estimated genome size. Since no assembly is representing 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. 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) (DeFraia *et al.*, 2010; Zampini *et al.*, 2015; Kleinboelting *et al.*, 2015; Pellaud *et al.*, 2018; Li *et al.*, 2018b; Wynn *et al.*, 2020; Pucker *et al.*, 2021) and Niederzenz-1 (Nd-1) (Pucker *et al.*, 2016) as well as several *Beta vulgaris* accessions (Dohm *et al.*, 2014; Tränkner *et al.*, 2016; Funk *et al.*, 2018) 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 (Lamesch *et al.*, 2012), AthNd-1_v2 (Pucker *et al.*, 2019), and RefBeet v1.5 (Dohm *et al.*, 2014; Holtgräwe *et al.*, 2017) 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, (Stadermann *et al.*, 2015)) and chondrome (BA000009.3, (Kubo *et al.*, 2000)) sequences were added to RefBeet v1.5 to allow proper placement of respective reads.

Genome sequences of *Brachypodium distachyon* strain Bd21 (GCF_000005505.3 (Vogel *et al.*, 2010)), *Solanum lycopersicum* (GCA_002954035.1 (Li *et al.*, 2018a)), *Vitis vinifera* cultivar Chardonnay (QGNW01000001.1 (Roach *et al.*, 2018)), *Oryza sativa* ssp. *japonica* cultivar Nipponbare (GCA_034140825.1 (Shang *et al.*, 2023)), and *Zea mays* cultivar DK105 (GCA_003709335.1 (Unterseer *et al.*, 2017)) were retrieved from the NCBI. Corresponding read data sets of *Brachypodium distachyon* ((Vogel *et al.*, 2010)), *Solanum lycopersicum* (Lang *et al.*, 2017; Li *et al.*, 2018a; Corem *et al.*, 2018; Yang *et al.*, 2019), *Vitis vinifera* (Roach *et al.*, 2018), *Oryza sativa* (Mahesh *et al.*, 2016; Shang *et al.*, 2023), and *Zea mays* ((Unterseer *et al.*, 2017)) were retrieved from the Sequence Read Archive (AdditionalFile1).

Genome size estimation

JellyFish2 v2.2.4 (Marçais & Kingsford, 2011) was applied for the generation of k-mer profiles which were subjected to GenomeScope (Vurture *et al.*, 2017). 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) (Liu *et al.*, 2020) and findGSE (Sun *et al.*, 2018) 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 is consequently a quality indicator.

Mapping-based genome size estimation

Despite some known biases (Grokhovsky *et al.*, 2011; van Heesch *et al.*, 2013; Poptsova *et al.*, 2014), 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.

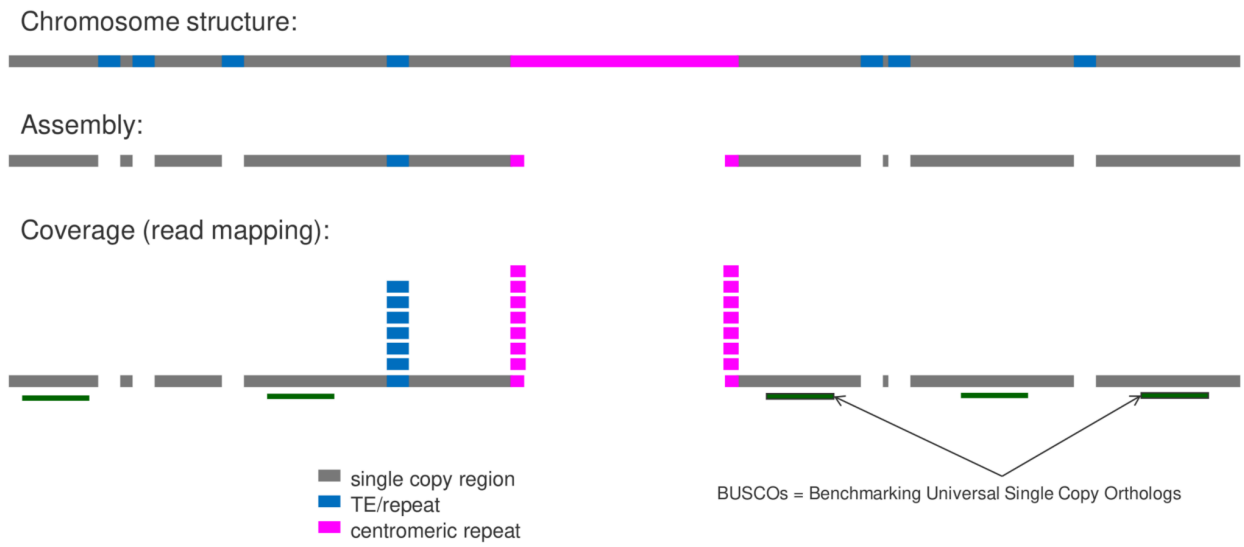


Figure 1: Concept diagram showing the working logic behind read mapping-based genome size estimation.

Underrepresented repeats and other regions like the unresolved paralogs of the SEC10 gene in the *Arabidopsis thaliana* genome sequence (Vukašinović *et al.*, 2014), 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) (Simão *et al.*, 2015; Manni *et al.*, 2021) 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.

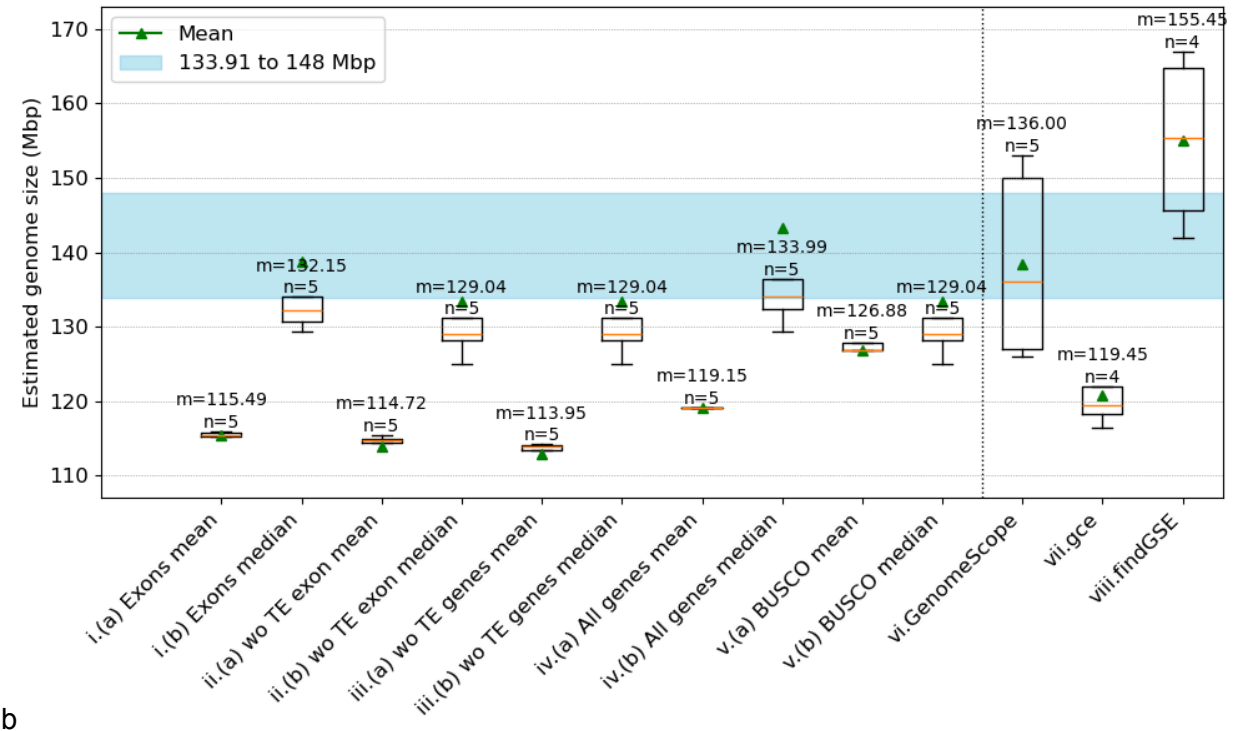
For short reads, BWA MEM v0.7 (Li, 2013) was applied for the read mapping and MarkDuplicates (Picard tools v2.14) (Broad Institute, 2019) was used to filter out reads originating from PCR duplicates. For long reads, minimap2 v2.24 (Li, 2018, 2021) was used for read mapping. Next, a previously described Python script (Pucker & Brockington, 2018) was deployed to generate coverage files, which provides information about the number of aligned sequencing reads covering each position of the reference sequence. Finally, MGSE v3.1 (https://github.com/bpucker/MGSE) was run on these coverage files to predict genome sizes independently for each data set.

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. 1). 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 GenomeScope, gce, and findGSE for k-mer sizes 19, 21, 23, and 25.

a


Mean
133.91 to 148 Mbp
Estimated genome size (Mbp)
170
160
150
140
130
120
110
m=115.49
n=5
m=132.15
n=5
m=114.72
n=5
m=129.04
n=5
m=113.95
n=5
m=129.04
n=5
m=119.15
n=5
m=133.99
n=5
m=126.88
n=5
m=129.04
n=5
m=136.00
n=5
m=119.45
n=4
m=155.45
n=4
i.(a) Exons mean
i.(b) Exons median
ii.(a) wo TE exon mean
ii.(b) wo TE exon median
iii.(a) wo TE genes mean
iii.(b) wo TE genes median
iv.(a) All genes mean
iv.(b) All genes median
v.(a) BUSCO mean
v.(b) BUSCO median
vi.GenomeScope
vii.gce
viii.findGSE

b

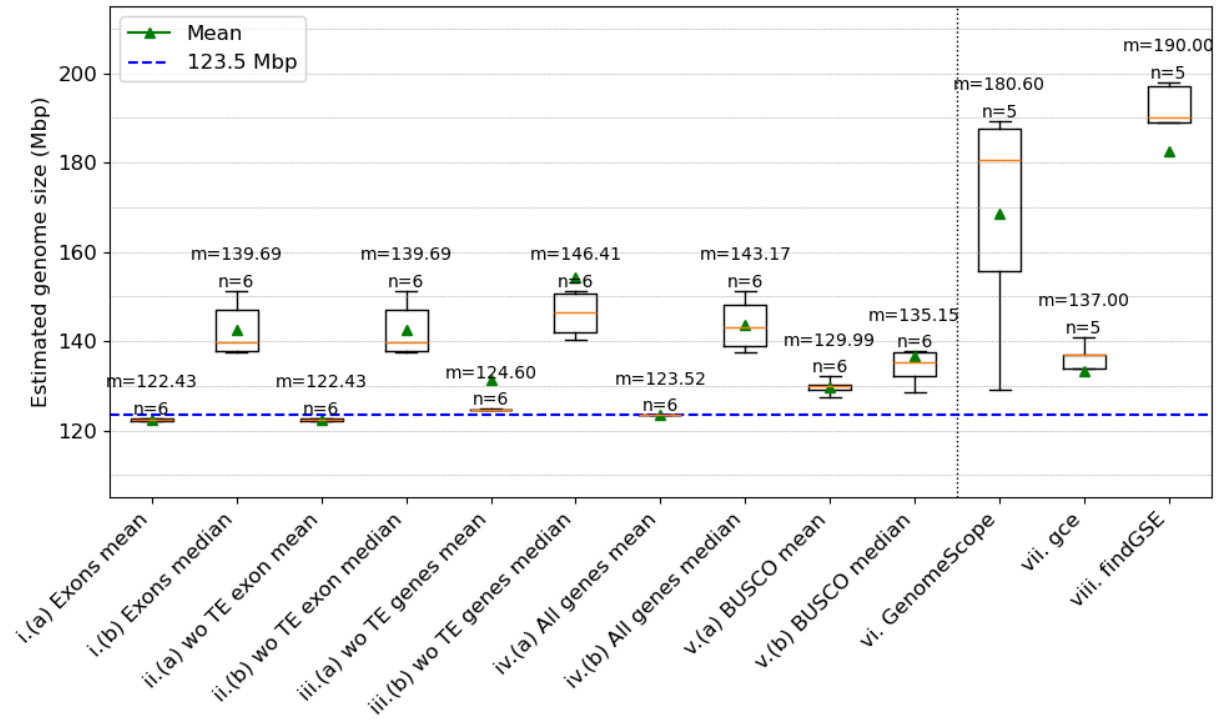

Mean
123.5 Mbp
Estimated genome size (Mbp)
200
180
160
140
120
m=122.43
n=6
m=139.69
n=6
m=122.43
n=6
m=139.69
n=6
m=124.60
n=6
m=146.41
n=6
m=123.52
n=6
m=143.17
n=6
m=129.99
n=6
m=135.15
n=6
m=180.60
n=5
m=137.00
n=5
m=190.00
n=5
i.(a) Exons mean
i.(b) Exons median
ii.(a) wo TE exon mean
ii.(b) wo TE exon median
iii.(a) wo TE genes mean
iii.(b) wo TE genes median
iv.(a) All genes mean
iv.(b) All genes median
v.(a) BUSCO mean
v.(b) BUSCO median
vi. GenomeScope
vii. gce
viii. findGSE

Fig. 1: 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, GenomeScope, 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 (green 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 (Hou *et al.*, 2022) as the lower boundary and one of the largest reported assembly sizes of *Arabidopsis thaliana* (Lian *et al.*, 2024) as the upper boundary (AdditionalFile6); 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* (Lian *et al.*, 2024) 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. GenomeScope 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. 1b). The reported assembly size of 123.5 Mbp is known to be an underestimation of the true genome size of the Nd-1 accession (Pucker *et al.*, 2019). 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 (Pucker *et al.*, 2016). Approximately 123.5 Mbp are assembled into pseudochromosomes which do not contain complete NORs or centromeric regions (Pucker *et al.*, 2019). 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 (Pucker *et al.*, 2016). Centromeric repeats which are only partially represented in the genome assembly (Pucker *et al.*, 2019) account for up to 11 Mbp (Pucker *et al.*, 2016). 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 GenomeScope and findGSE are very unlikely, as they contradict most estimations of *A. thaliana* genome sizes (Arumuganathan & Earle, 1991; Pucker *et al.*, 2016, 2019; Sun *et al.*, 2018).

MGSE was also used to estimate the genome size of the Col-0 accession using long read data sets of 14 GABI-Kat lines (Pucker *et al.*, 2021) (Fig. 2, AdditionalFile3). Since BUSCO estimates were chosen to be better from the previous *A. thaliana* analyses, BUSCO based estimations were done for all these lines.

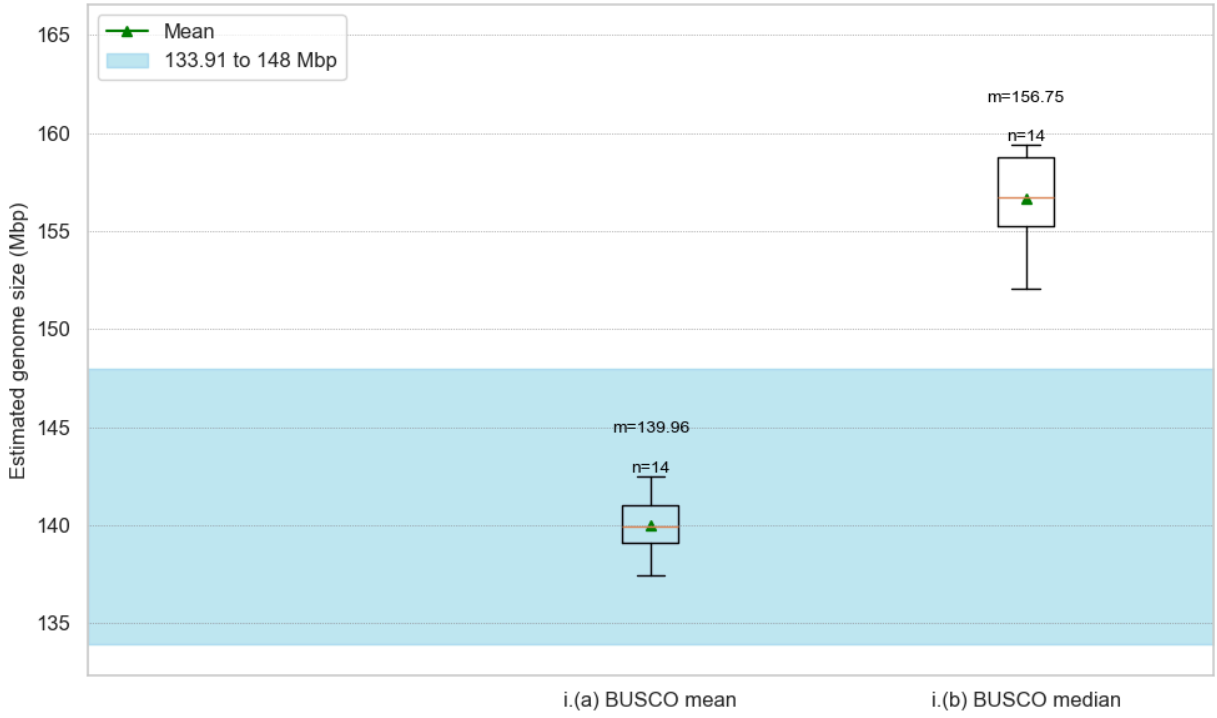


Fig. 2: MGSE estimations using long read sequencing data sets of 14 GABI-Kat lines (Col-0 accession). The blue region shows the expected genome size range with the near complete assembly size of Col-0 (Hou *et al.*, 2022) as the lower boundary and one of the largest reported assembly sizes of *Arabidopsis thaliana* (Lian *et al.*, 2024) as the upper boundary (AdditionalFile6).

The MGSE BUSCO mean fits well into the expected genome size region of Col-0, while the median estimate is slightly higher than the largest reported assembly size considered here. Potential reason for this deviation could be attributed to the presence of a high number of InDels in long reads. 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 other kmer-based approaches, making it valuable at a time where 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. 3, AdditionalFile4) which were analyzed by re-sequencing as part of the 1001 genome project (Alonso-Blanco *et al.*, 2016). 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. This comparison revealed systematic differences between findGSE, gce, and GenomeScope with respect to the average predicted genome size. findGSE tends to predict larger genome sizes than gce and GenomeScope. Very large genome sizes could have biological explanations like polyploidization events.

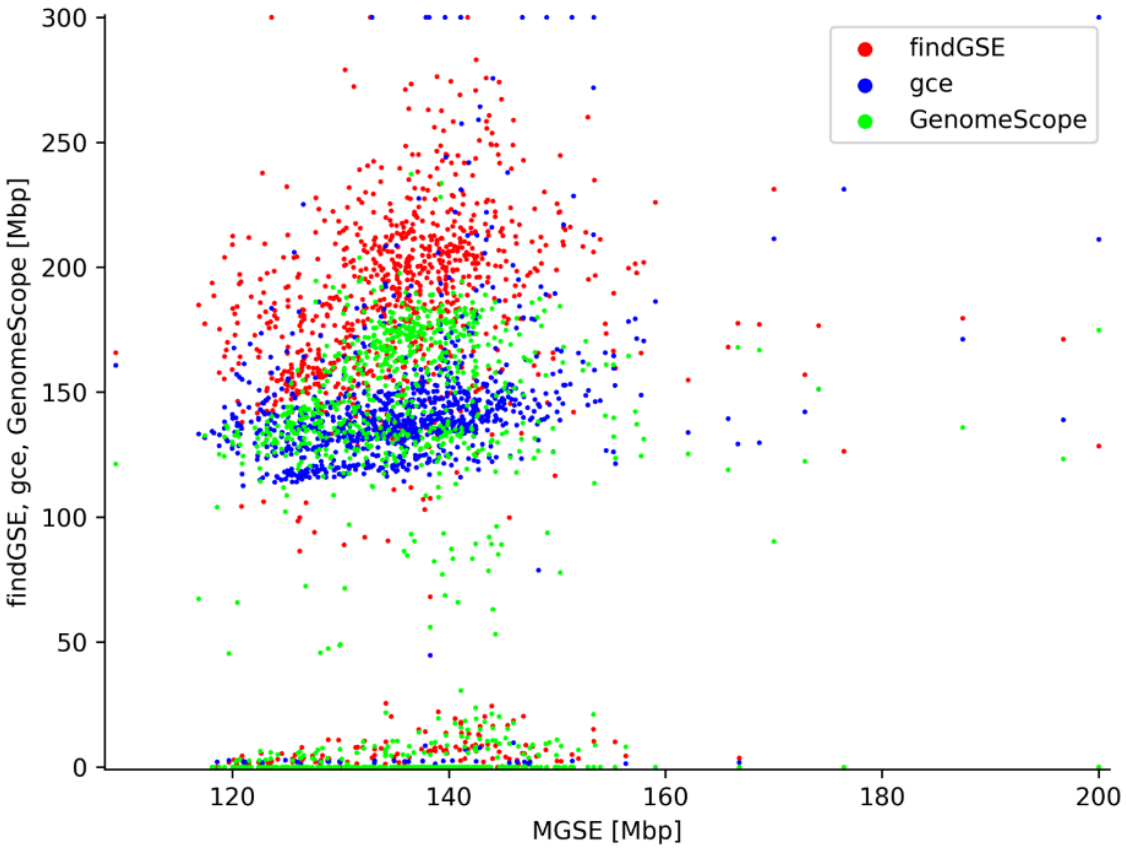


Fig. 3: Genome size estimations of *Arabidopsis thaliana* accessions. MGSE, findGSE, gce, and GenomeScope were deployed to predict the genome sizes of 1,028 *A. thaliana* accessions based on sequence read data sets (AdditionalFile4). Extreme outliers above 200 Mbp (MGSE) or 300 Mbp (other tools) are 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, GenomeScope, gce, and findGSE to assess the applicability to larger and more complex genomes. 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 (Dohm *et al.*, 2014; Holtgräwe *et al.*, 2017), 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 (Kowar *et al.*, 2016). The largest reported assembly of the reference accession KWS2320 has a size of 573 Mbp (Sielemann *et al.*, 2023) 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 (Sielemann *et al.*, 2023). 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 (Arumuganathan & Earle, 1991) and 731 Mbp (Dohm *et al.*, 2014). 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 (Arumuganathan & Earle, 1991; Dohm *et al.*, 2014). It is noticeable that the findGSE analysis failed for a number of samples. It is also important to note that GenomeScope 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 found to be optimal for the *B. vulgaris* analyses. 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.

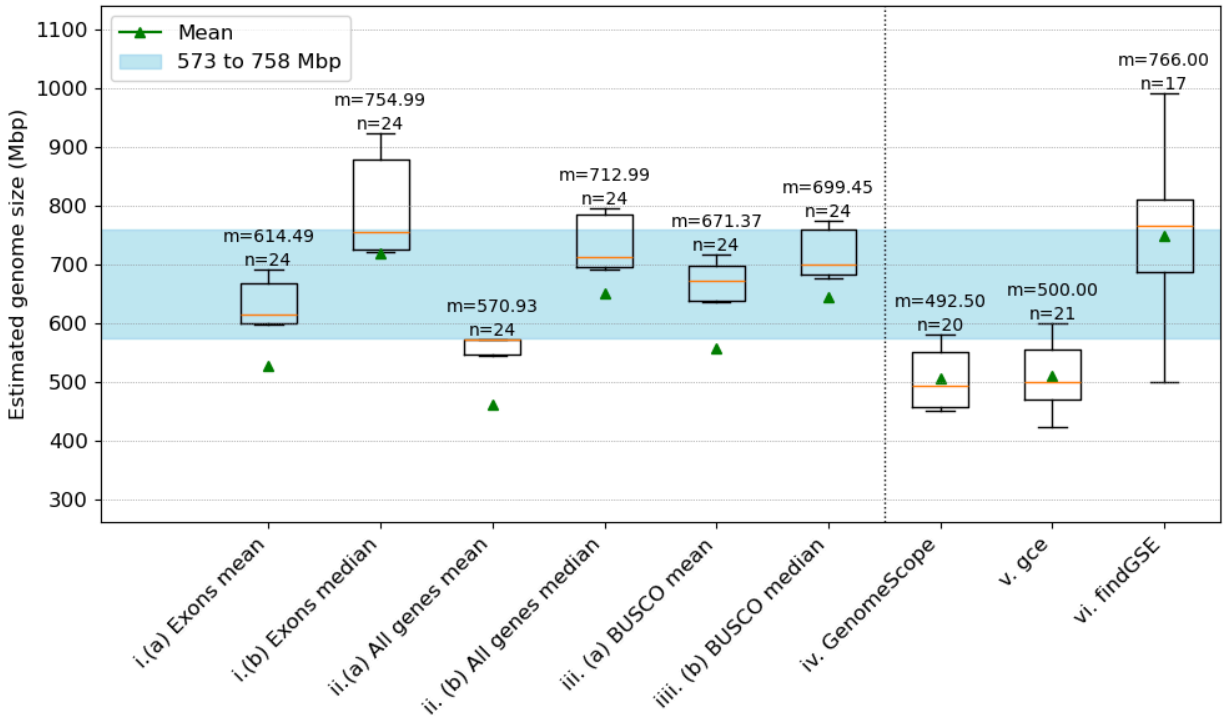


Fig. 4: Comparison of *Beta vulgaris* genome size estimations. The genome size of *B. vulgaris* was predicted by MGSE, GenomeScope, gce, and findGSE. Different MGSE approaches were evaluated differing by the set of regions for the average coverage calculation (e.g. all genes) and the methods for the 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 (green triangles). The blue region shows the expected genome size range with the largest reported assembly for *Beta vulgaris* (Sielemann *et al.*, 2023) as the lower boundary and the largest estimated genome size of *Beta vulgaris* (Arumuganathan & Earle, 1991) as the upper boundary (AdditionalFile6).

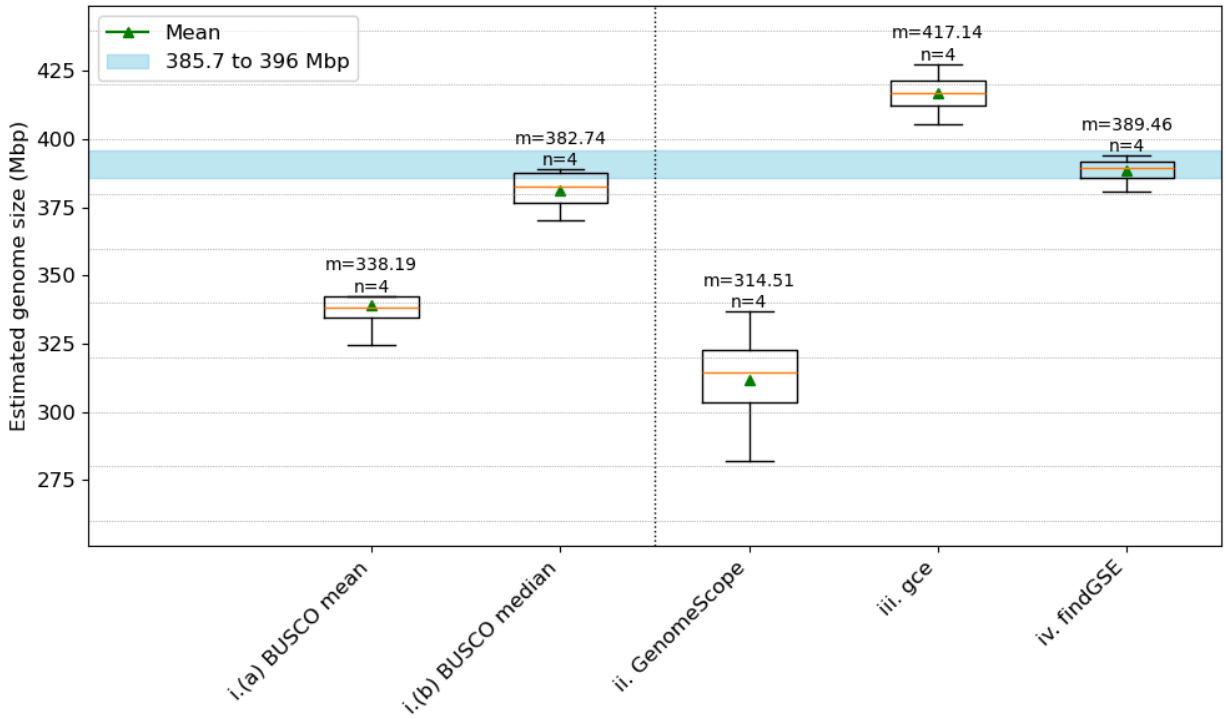
Oryza sativa genome size

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 (Huang *et al.*, 2021). Recently, a complete genome sequence of *Oryza sativa* ssp. japonica cv. Nipponbare was reported with 385.7 Mbp (Shang *et al.*, 2023). It was obtained using a hybrid assembly approach combining reads from illumina, ONT and PacBio sequencing technologies with a Hi-C dataset (Shang *et al.*, 2023).

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 GenomeScope 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. 5a. The comparative estimations by MGSE using short and long reads are shown in Fig. 5b.

a



b

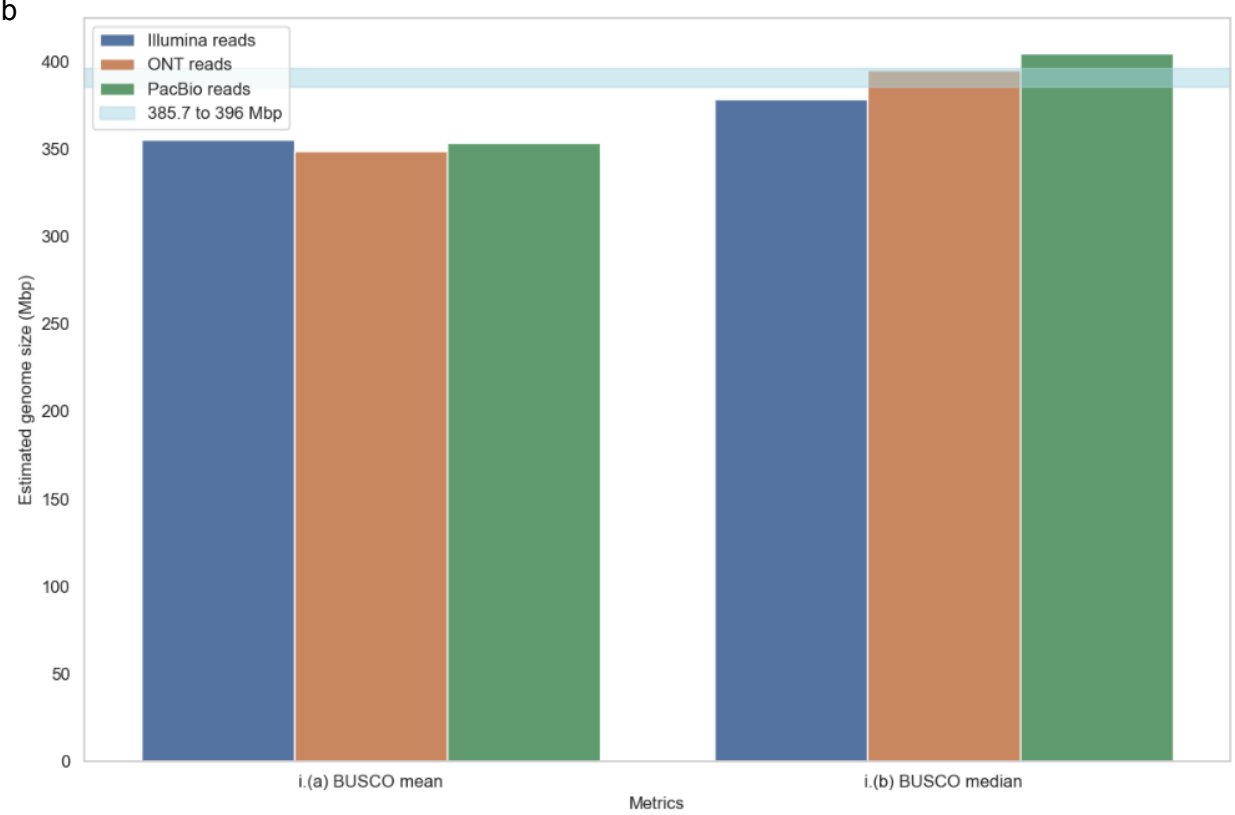


Fig. 5: (a) Comparison of *Oryza sativa* genome size estimations using short reads. (b) Comparison of MGSE genome size estimations on short and long reads of *O. sativa.* The blue region shows the expected genome size values (AdditionalFile6). It has the size of the near complete genome sequence of *O. sativa* as the lower boundary and the size of the largest reported assembly of *O. sativa* as the upper boundary (Song *et al.*, 2021).

While MGSE BUSCO median estimate and findGSE offer size estimates very close to the size of the presumable complete genome sequence, BUSCO mean estimates of MGSE and GenomeScope underestimate the genome size. gce slightly overestimates the size, but the estimate seems reliable and not grossly predicted.

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 accurate 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, and *Vitis vinifera* was selected due to a very high heterozygosity. The predictions of MGSE are generally in the same range as the predictions generated by GenomeScope, gce, and findGSE (AdditionalFile5, AdditionalFile7, AdditionalFile8, AdditionalFile9, and AdditionalFile10). 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. GenomeScope and gce predict genome sizes below the assembly size (AdditionalFile7). The *Z. mays* genome size is underestimated by all four tools. However, MGSE outperforms GenomeScope 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. GenomeScope underestimated the genome size of *V. vinifera* as evident from the assembly size exceeding the estimation (AdditionalFile10).

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 reads can cover contig ends. The shorter a

contig, the more the apparent coverage of this contig is reduced. 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 part of the assembly process (Jiao *et al.*, 2017; Michael *et al.*, 2018; Hibrand Saint-Oyant *et al.*, 2018; Pucker *et al.*, 2019). Future genome sequencing projects are likely to generate high continuity assemblies (Pucker *et al.*, 2022) 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 (Kumar & Blaxter, 2011; Salter *et al.*, 2014; Strong *et al.*, 2014; Pucker *et al.*, 2016; Mallet *et al.*, 2017).

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.

All methods are affected by DNA of the plastome and chondrome integrated into the nuclear chromosomes (Ayliffe *et al.*, 1998; Michalovova *et al.*, 2013). 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 (Dohm *et al.*, 2008; Ross *et al.*, 2013). Both, extremely GC-rich and AT-rich fragments, respectively, are underrepresented in the sequencing output mainly due to biases introduced by PCR (Aird *et al.*, 2011; Benjamini & Speed, 2012). Sophisticated methods were developed to correct coverage values based on the GC content of the underlying

sequence (Benjamini & Speed, 2012; Love *et al.*, 2016; Teng & Irizarry, 2017). 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 (Pucker & Brockington, 2018). 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 and cannot be inferred from the reads due to the above mentioned sequencing bias. 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 plant species. Several of these genome sequences should be very close to a perfect representation of the genome 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. Although not evaluated in this study, we envision that MGSE could be generally applied 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 GenomeScope.

AdditionalFile5: *B. vulgaris, Zea mays, Oryza sativa, Brachypodium distachyon, Solanum lycopersicum*, and *Vitis vinifera* genome size prediction values for all different approaches.

AdditionalFile6: List of near complete/ complete and largest reported genome assembly sizes for the different organisms analyzed in this study

AdditionalFile7: Genome size estimation of *Brachypodium distachyon*.

AdditionalFile8: Genome size estimation of *Zea mays*.

AdditionalFile9: Genome size estimation of *Solanum lycopersicum*.

AdditionalFile10: Genome size estimation of *Vitis vinifera*.

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

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. **2011**. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**: R18. doi: 10.1186/gb-2011-12-2-r18.

Albini S m.1994. A karyotype of the Arabidopsis thaliana genome derived from synaptonemal complex analysis at prophase I of meiosis. *The Plant Journal* **5**: 665–672. doi: 10.1111/j.1365-313X.1994.00665.x.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, *et al.*2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**: 481–491. doi: 10.1016/j.cell.2016.05.063.

Arumuganathan K, Earle ED. **1991**. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* **9**: 208–218. doi: 10.1007/BF02672069.

Ayliffe MA, Scott NS, Timmis JN. **1998**. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. *Molecular Biology and Evolution* **15**: 738–745. doi: 10.1093/oxfordjournals.molbev.a025977.

Benjamini Y, Speed TP. **2012**. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**: e72. doi: 10.1093/nar/gks001.

Bennett MD, Leitch IJ. **2011**. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany* **107**: 467–590. doi: 10.1093/aob/mcq258.

Bennett MD, Leitch IJ, Price HJ, Johnston JS. **2003**. 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. *Annals of Botany* **91**: 547–557. doi: 10.1093/aob/mcg057.

Bennett MD, Smith JB. **1991**. Nuclear DNA Amounts in Angiosperms. *Philosophical Transactions: Biological Sciences* **334**: 309–345.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. **2015**. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* **33**: 623–630. doi: 10.1038/nbt.3238.

Bevan M, Walsh S. **2005**. The Arabidopsis genome: a foundation for plant research. *Genome Research* **15**: 1632–1642. doi: 10.1101/gr.3723405.

Broad Institute. **2019**. Picard Tools.

Chikhi R, Medvedev P. **2014**. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**: 31–37. doi: 10.1093/bioinformatics/btt310.

Corem S, Doron-Faigenboim A, Jouffroy O, Maumus F, Arazi T, Bouché N. **2018**. Redistribution of CHH Methylation and Small Interfering RNAs across the Genome of Tomato ddm1 Mutants. *The Plant Cell* **30**: 1628–1644. doi: 10.1105/tpc.18.00167.

DeFraia CT, Zhang X, Mou Z. **2010**. Elongator subunit 2 is an accelerator of immune responses in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology* **64**: 511–523. doi: 10.1111/j.1365-313X.2010.04345.x.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. **2008**. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**: e105. doi: 10.1093/nar/gkn425.

Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R, *et al.*2014. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature* **505**: 546–549. doi: 10.1038/nature12817.

Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R. **1998**. Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. *Annals of Botany* **82**: 17–26. doi: 10.1093/oxfordjournals.aob.a010312.

Francis DM, Hulbert SH, Michelmore RW. **1990**. Genome size and complexity of the obligate

fungal pathogen, Bremia lactucae. *Experimental Mycology* **14**: 299–309. doi: 10.1016/0147-5975(90)90053-V.
Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I. **2002**. Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 14584–14589. doi: 10.1073/pnas.212325299.
Funk A, Galewski P, McGrath JM. **2018**. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *The Plant Journal* **95**: 659–671. doi: 10.1111/tpj.13977.
Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, *et al.*2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108**: 1513–1518. doi: 10.1073/pnas.1017351108.
Goodwin S, McPherson JD, McCombie WR. **2016**. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**: 333–351. doi: 10.1038/nrg.2016.49.
Grokhovsky SL, Il'icheva IA, Nechipurenko DYu, Golovkin MV, Panchenko LA, Polozov RV, Nechipurenko YD. **2011**. Sequence-Specific Ultrasonic Cleavage of DNA. *Biophysical Journal* **100**: 117–125. doi: 10.1016/j.bpj.2010.10.052.
van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, Toonen P, de Bruijn E, Shull JD, Aitman TJ, Cuppen E, *et al.*2013. Systematic biases in DNA copy number originate from isolation procedures. *Genome Biology* **14**: R33. doi: 10.1186/gb-2013-14-4-r33.
Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, Bourke PM, Daccord N, Leus L, Schulz D, *et al.*2018. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. *Nature Plants* **4**: 473–484. doi: 10.1038/s41477-018-0166-1.
Holtgräwe D, Rosleff Sörensen T, Parol-Kryger R, Pucker B, Viehöver P, Kleinbölting N, Weisshaar B. **2017**. Low coverage re-sequencing in Beta vulgaris (sugar beet) for anchoring assembly sequences to genomic position.
Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. **2022**. A near-complete assembly of an Arabidopsis thaliana genome. *Molecular Plant* **15**: 1247–1250. doi: 10.1016/j.molp.2022.05.014.
Huang C, Chen Z, Liang C. **2021**. *Oryza* pan-genomics: A new foundation for future rice research and improvement. *The Crop Journal* **9**: 622–632. doi: 10.1016/j.cj.2021.04.003.
Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, *et al.*2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research* **27**: 778–786. doi: 10.1101/gr.213652.116.
Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J, *et al.*2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data* **1**: 140045. doi: 10.1038/sdata.2014.45.
Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B. **2015**. The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. *Molecular Plant* **8**: 1651–1664. doi: 10.1016/j.molp.2015.08.011.
Kowar T, Zakrzewski F, Macas J, Kobližková A, Viehoever P, Weisshaar B, Schmidt T. **2016**. Repeat Composition of CenH3-chromatin and H3K9me2-marked heterochromatin in Sugar Beet (Beta vulgaris). *BMC Plant Biology* **16**: 120. doi: 10.1186/s12870-016-0805-5.
Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T. **2000**. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNACys(GCA). *Nucleic Acids Research* **28**: 2571–2576.
Kumar A, Bennetzen JL. **1999**. Plant retrotransposons. *Annual Review of Genetics* **33**: 479–532. doi: 10.1146/annurev.genet.33.1.479.

Kumar S, Blaxter ML. **2011**. Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis (Philadelphia, Pa.)* **55**: 119–126. doi: 10.1007/s13199-012-0154-6.
Kurtz S, Narechania A, Stein JC, Ware D. **2008**. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**: 517. doi: 10.1186/1471-2164-9-517.
Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.*2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**: D1202–D1210. doi: 10.1093/nar/gkr1090.
Lang Z, Wang Y, Tang K, Tang D, Datsenka T, Cheng J, Zhang Y, Handa AK, Zhu J-K. **2017**. 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* **114**: E4511–E4519. doi: 10.1073/pnas.1705233114.
Leutwiler LS, Hough-Evans BR, Meyerowitz EM. **1984**. The DNA of Arabidopsis thaliana. *Molecular and General Genetics MGG* **194**: 15–23. doi: 10.1007/BF00383491.
Li H. **2013**. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*.
Li H. **2018**. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. doi: 10.1093/bioinformatics/bty191.
Li H. **2021**. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572–4574. doi: 10.1093/bioinformatics/btab705.
Li J, Chitwood J, Menda N, Mueller L, Hutton SF. **2018a**. Linkage between the I-3 gene for resistance to Fusarium wilt race 3 and increased sensitivity to bacterial spot in tomato. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **131**: 145–155. doi: 10.1007/s00122-017-2991-4.
Li J, Liang W, Li Y, Qian W. **2018b**. APURINIC/APYRIMIDINIC ENDONUCLEASE2 and ZINC FINGER DNA 3′-PHOSPHOESTERASE Play Overlapping Roles in the Maintenance of Epigenome and Genome Stability. *The Plant Cell* **30**: 1954–1970. doi: 10.1105/tpc.18.00287.
Li X, Waterman MS. **2003**. Estimating the Repeat Structure and Length of DNA Sequences Using ℓ-Tuples. *Genome Research* **13**: 1916–1922. doi: 10.1101/gr.1251803.
Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, *et al.*2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**: 265–272. doi: 10.1101/gr.097261.109.
Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R. **2024**. A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics* **56**: 982–991. doi: 10.1038/s41588-024-01715-9.
Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. **2020**. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. doi: 10.48550/arXiv.1308.2012.
Love MI, Hogenesch JB, Irizarry RA. **2016**. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology* **34**: 1287–1291. doi: 10.1038/nbt.3682.
Mahesh HB, Shirke MD, Singh S, Rajamani A, Hittalmani S, Wang G-L, Gowda M. **2016**. Indica rice genome assembly, annotation and mining of blast disease resistance genes. *BMC Genomics* **17**: 242. doi: 10.1186/s12864-016-2523-7.
Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H. **2017**. PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. *Bioinformatics* **33**: 3283–3285. doi: 10.1093/bioinformatics/btx396.
Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. **2021**. 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* **38**: 4647–4654. doi: 10.1093/molbev/msab199.
Marçais G, 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.
Mardis ER. **2017**. DNA sequencing technologies: 2006–2016. *Nature Protocols* **12**: 213–218. doi: 10.1038/nprot.2016.182.
Metzker ML. **2010**. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**: 31–46. doi: 10.1038/nrg2626.
Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. **2018**. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications* **9**: 541. doi: 10.1038/s41467-018-03016-2.
Michalovova M, Vyskot B, Kejnovsky E. **2013**. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity* **111**: 314–320. doi: 10.1038/hdy.2013.51.
Pellaud S, Bory A, Chabert V, Romanens J, Chaisse-Leal L, Doan AV, Frey L, Gust A, Fromm KM, Mène-Saffrané L. **2018**. WRINKLED1 and ACYL-COA:DIACYLGLYCEROL ACYLTRANSFERASE1 regulate tocochromanol metabolism in Arabidopsis. *The New Phytologist* **217**: 245–260. doi: 10.1111/nph.14856.
Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. **2020**. Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3 Genes|Genomes|Genetics* **10**: 3047–3060. doi: 10.1534/g3.120.401028.
Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovsky SL. **2014**. Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports* **4**: 4532. doi: 10.1038/srep04532.
Pucker B, Brockington SF. **2018**. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics* **19**: 980. doi: 10.1186/s12864-018-5360-z.
Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. **2016**. A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. *PLOS ONE* **11**: e0164321. doi: 10.1371/journal.pone.0164321.
Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B. **2019**. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLOS ONE* **14**: e0216233. doi: 10.1371/journal.pone.0216233.
Pucker B, Irisarri I, Vries J de, Xu B. **2022**. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology* **3**: e5. doi: 10.1017/qpb.2021.18.
Pucker B, Kleinbölting N, Weisshaar B. **2021**. Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics* **22**: 599. doi: 10.1186/s12864-021-07877-8.
Roach MJ, Johnson DL, Bohlmann J, Vuuren HJJ van, Jones SJM, Pretorius IS, Schmidt SA, Borneman AR. **2018**. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLOS Genetics* **14**: e1007807. doi: 10.1371/journal.pgen.1007807.
Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. **2013**. Characterizing and measuring bias in sequence data. *Genome Biology* **14**: R51. doi: 10.1186/gb-2013-14-5-r51.
Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. **2014**. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**: 87. doi: 10.1186/s12915-014-0087-z.
Schmuths H, Meister A, Horres R, Bachmann K. **2004**. Genome Size Variation among

Accessions of Arabidopsis thaliana. *Annals of Botany* **93**: 317–321. doi: 10.1093/aob/mch037.
Shang L, He W, Wang T, Yang Y, Xu Q, Zhao X, Yang L, Zhang H, Li X, Lv Y, *et al.*2023. A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* **16**: 1232–1236. doi: 10.1016/j.molp.2023.08.003.
Sielemann K, Pucker B, Orsini E, Elashry A, Schulte L, Viehöver P, Müller AE, Schechert A, Weisshaar B, Holtgräwe D. **2023**. Genomic characterization of a nematode tolerance locus in sugar beet. *BMC Genomics* **24**: 748. doi: 10.1186/s12864-023-09823-2.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. **2015**. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212. doi: 10.1093/bioinformatics/btv351.
Sivadasan N, Srinivasan R, Goyal K. **2016**. Kmerlight: fast and accurate k-mer abundance estimation. doi: 10.48550/arXiv.1609.05626.
Song J-M, Xie W-Z, Wang S, Guo Y-X, Koo D-H, Kudrna D, Gong C, Huang Y, Feng J-W, Zhang W, *et al.*2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant* **14**: 1757–1767. doi: 10.1016/j.molp.2021.06.018.
Stadermann KB, Weisshaar B, Holtgräwe D. **2015**. SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. *BMC Bioinformatics* **16**: 295. doi: 10.1186/s12859-015-0726-6.
Strong MJ, Xu G, Morici L, Bon-Durant SS, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK. **2014**. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLOS Pathogens* **10**: e1004437. doi: 10.1371/journal.ppat.1004437.
Sun H, Ding J, Piednoël M, Schneeberger K. **2018**. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics (Oxford, England)* **34**: 550–557. doi: 10.1093/bioinformatics/btx637.
Teng M, Irizarry RA. **2017**. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. *Genome Research* **27**: 1930–1938. doi: 10.1101/gr.220673.117.
The Arabidopsis Genome Initiative. **2000**. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796–815. doi: 10.1038/35048692.
Tränkner C, Lemnian IM, Emrani N, Pfeiffer N, Tiwari SP, Kopisch-Obuch FJ, Vogt SH, Müller AE, Schilhabel M, Jung C, *et al.*2016. 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* **7**.
Unterseer S, Seidel MA, Bauer E, Haberer G, Hochholdinger F, Opitz N, Marcon C, Baruch K, Spannagl M, Mayer KFX, *et al.*2017. European Flint reference sequences complement the maize pan-genome. : 103747. doi: 10.1101/103747.
Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, *et al.*2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**: 763–768. doi: 10.1038/nature08747.
Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, Synek L. **2014**. Dissecting a Hidden Gene Duplication: The Arabidopsis thaliana SEC10 Locus. *PLOS ONE* **9**: e94077. doi: 10.1371/journal.pone.0094077.
Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. **2017**. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202–2204. doi: 10.1093/bioinformatics/btx153.
Wynn E, Purfeerst E, Christensen A. **2020**. Mitochondrial DNA Repair in an Arabidopsis thaliana Uracil N-Glycosylase Mutant. *Plants* **9**: 261. doi: 10.3390/plants9020261.
Yang Y, Tang K, Datsenka TU, Liu W, Lv S, Lang Z, Wang X, Gao J, Wang W, Nie W, *et al.*2019. Critical function of DNA methyltransferase 1 in tomato development and regulation of the DNA methylome and transcriptome. *Journal of Integrative Plant Biology* **61**: 1224–1242. doi: 10.1111/jipb.12778.

Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. **2015**. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in Arabidopsis and humans. *Genome Research* **25**: 645–654. doi: 10.1101/gr.188573.114.
Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, Patel V, Velikkakam James G, Koornneef M, Ossowski S, *et al.*2016. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences* **113**: E4052–E4060. doi: 10.1073/pnas.1607532113.