Mapping-based genome size estimation

Boas Pucker[1,2*]

1 Genetics and Genomics of Plants, Bielefeld University, Bielefeld, Germany

2 Center for Biotechnology (CeBiTec); Bielefeld University, Bielefeld, Germany

Email: bpucker@cebitec.uni-bielefeld.de

ORCID: 0000-0002-3321-7471

* corresponding author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

Key words: NGS, genome sequencing, k-mer profile, comparative genomics, *Arabidopsis thaliana*, *Beta vulgaris, Solanum lycopersicum, Brachypodium distachyon,Vitis vinifera, Zea mays*

Abstract

While the size of chromosomes can be measured under a microscope, the size of genomes cannot be measured precisely. Biochemical methods and k-mer distribution-based approaches allow only estimations. An alternative approach to predict the genome size based on high contiguity assemblies and short read mappings is presented here and optimized on *Arabidopsis thaliana* and *Beta vulgaris*. *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 [1], the combined length of all DNA molecules in a single cell is still unknown. Almost 20 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. Next, genome sizes can be estimated based on unique k-mers or a complete k-mer profile. Many assemblers like SOAPdenovo [16] and ALLPATHS-LG [17] perform an internal estimation of the genome size to infer an expected assembly size. Recently, dedicated tools for the genome size estimation like GenomeScope [18] and findGSE [19] were developed. Although the authors considered and addressed a plethora of issues with real data [18], 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* [19, 20], 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 [21, 22]. Nevertheless, some highly repetitive regions of plant genomes like nucleolus organizing region (NOR) and centromeres remain usually unassembled [20, 23, 24]. 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 [25] have been proven to be inaccurate based on insights from recent assemblies [20, 24, 26–28]. However, various methods still predict genome sizes between 125 Mbp and 165 Mbp for diploid *A. thaliana* accessions [26, 29–31]. Substantial technical variation is observed not only between methods, but also between different labs or instruments [32]. 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 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.

Methods

Data sets

Sequencing data sets of the *A. thaliana* accessions Columbia-0 (Col-0) [33–38] and Niederzenz-1 (Nd-1) [31] as well as several *Beta vulgaris* accessions [39–41] were retrieved from the Sequence Read Archive (AdditionalFile 1). Only the paired-end fraction of the two included Nd-1 mate pair libraries was included in this analysis. Genome assembly versions TAIR9 [42], AthNd-1_v1 [31], AthNd-1_v2 [24], and RefBeet v1.5 [39, 43] served as references in the read mapping process. The *A. thaliana* assemblies, TAIR9 and Ath-Nd-1_v2, already included plastome and chondrome sequences. These subgenome sequences of Ath-Nd-1_v2 were added to Ath-Nd-1_v1 as this assembly was previously cleaned of such sequences. Plastome (KR230391.1, [44]) and chondrome (BA000009.3, [45]) sequences were added to RefBeet v1.5 to allow proper placement of respective reads.

Genome sequences of *Brachypodium distachyon* strain Bd21 (GCF_000005505.3 [46]), *Solanum lycopersicum* (GCA_002954035.1 [47]), *Vitis vinifera* cultivar Chardonnay (QGNW01000001.1 [48]), and *Zea mays* cultivar DK105 (GCA_003709335.1 [49]) were retrieved from the NCBI. Corresponding read data sets were retrieved from the Sequence Read Archive (AdditionalFile1).

Genome size estimation

JellyFish2 v2.2.4 [11] was applied for the generation of k-mer profiles which were subjected to GenomeScope [18]. 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 [19] 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 [50–52], 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: $N = L / C$. Underrepresented repeats and other regions 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 values i.e. plastome and chondrome sequences are included during the mapping phase to allow correct placement of reads, but are excluded from MGSE. 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) [53] 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. GFF 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.

BWA MEM v0.7 [54] was applied for the read mapping and MarkDuplicates (Picard tools v2.14) [55] was used to filter out reads originating from PCR duplicates. Next, a previously described Python script [56] was deployed to generate coverage files, which provide information about the number of aligned sequencing reads covering each position of the reference sequence. Finally, MGSE (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.

Many estimations of the Col-0 genome size are below the assembly size of 120 Mbp [26] and display substantial variation between samples (Fig. 1a). Due to low variation between different samples and a likely average genome size the BUSCO-based approaches appeared promising. GenomeScope predicted a similar genome size, while gce reported consistently much smaller values. findGSE predicted on average a substantially larger genome size. Final sample sizes below six indicated that prediction processes failed e.g. due to insufficient read numbers.

The variation among the estimated genome sizes of Nd-1 was smaller than the variation between the Col-0 samples (Fig. 1). BUSCO-based estimations differed substantially between mean and median with respect to the variation between samples (Fig. 1b). Therefore, the average coverage is probably more reliably calculated via mean than via median. While gce predicted as reasonable genome size for Nd-1, the average predictions by GenomeScope and findGSE are very unlikely, as they contradict most estimations of *A. thaliana* genome sizes [6, 19, 24, 31].

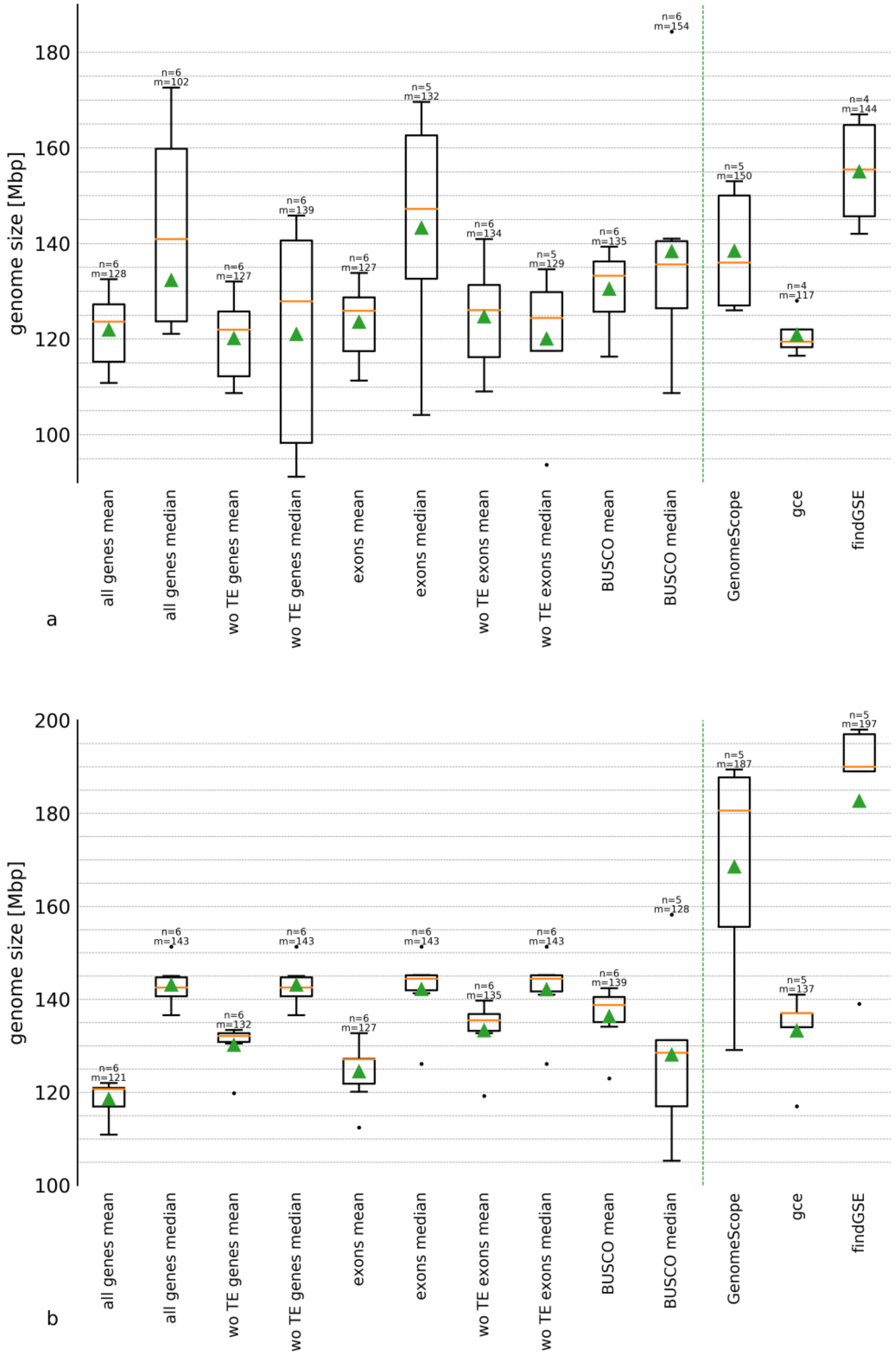

Fig. 1: Comparison of *Arabidopsis thaliana* genome size estimations.

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 genome size estimation of about 139 Mbp inferred for Nd-1 through integration of all analyses is slightly below previous estimations of about 146 Mbp [31]. Approximately 123.5 Mbp are assembled into pseudochromosomes which do not contain complete NORs or centromeric regions [24]. 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 [31]. Centrometric repeats which are only partially represented in the genome assembly [24] account for up to 11 Mbp [31]. In summary, the Nd-1 genome size is expected to be around 138-140 Mbp. The BUSCOs which occur actually with a single copy in Ath-Nd1_v2 emerged as the best set of reference regions for MGSE.

The relevance of very high assembly contiguity was assessed by comparing results of AthNd-1_v1 (AdditionalFile3), which is based on short Illumina reads, to results of AthNd-1_v2 (AdditionalFile2), which is based on long Single Molecule Real Time sequencing (PacBio) reads. The genome size predictions based on AthNd-1_v2 were substantially more accurate. Reads are not mapped to the ends of contigs or scaffolds. This has only a minor influence on large contigs, because a few small regions at the ends with lower coverage can be neglected. However, the average coverage of smaller contigs might be biased as the relative contribution of contig ends weights stronger. In addition, the representation of centrometric repeats and transposable elements increases with higher assembly size and contiguity [24].

The feasibility of MGSE was further demonstrated by estimating the genome sizes of 1,028 *A. thaliana* accessions (Fig. 2, AdditionalFile4) which were analyzed by re-sequencing as part of the 1001 genome project [57]. Most predictions by MGSE 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.

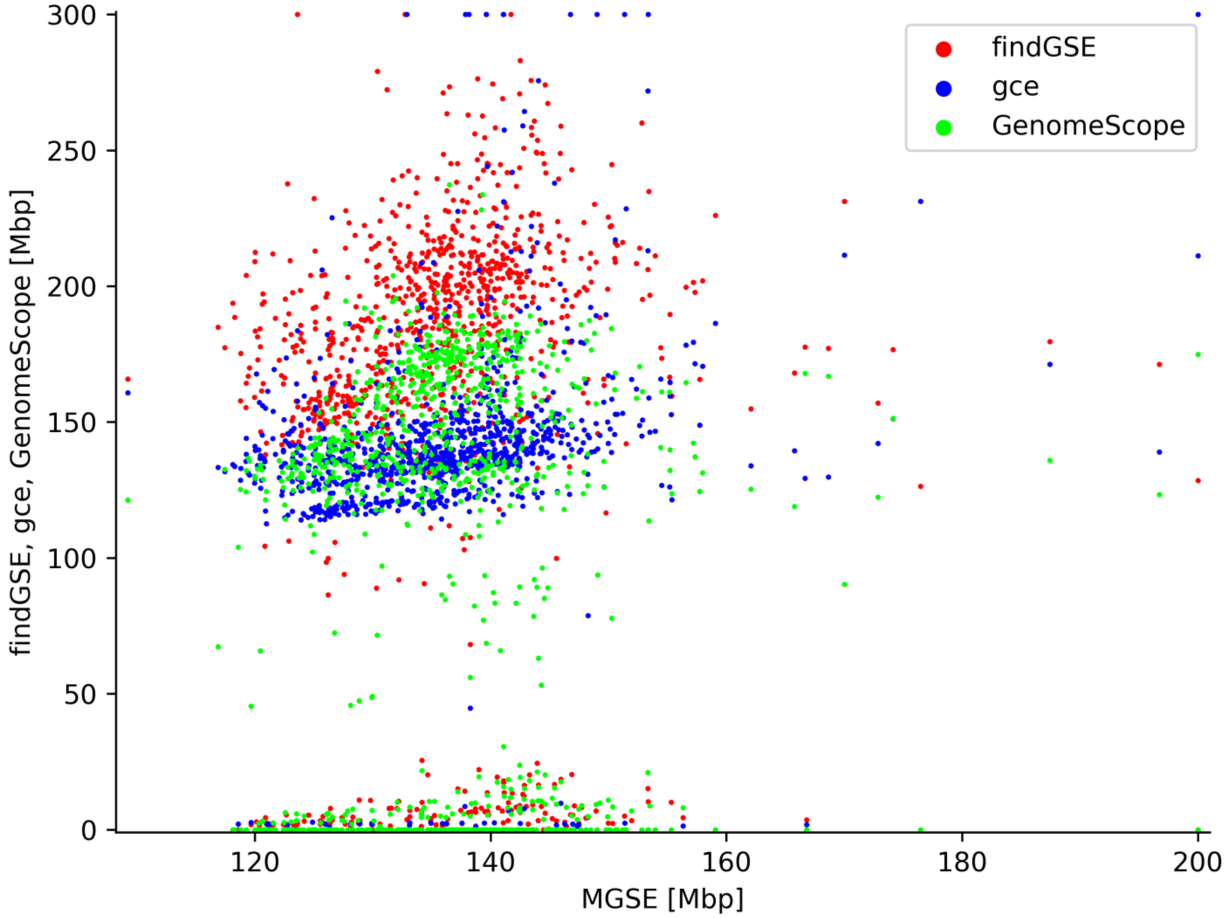

Fig. 2: Genome size estimations of *Arabidopsis thaliana* accessions.

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 (Fig. 3, 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. Since the current RefBeet v1.5 assembly represents 567 Mbp [39, 43] of the genome, all estimations below this value can be discarded as erroneous. Therefore, the mean-based approaches relying on all genes or just the BUSCOs as reference region for the sequencing coverage estimation outperformed all other approaches (Fig. 3). When comparing the *A. thaliana* and *B. vulgaris* analyses, the calculation of an average coverage in all BUSCOs, which are actually present as a single copy in the investigated genome, appears to be the most promising approach. While GenomeScope and gce underestimate the genome size, the predictions by findGSE are extremely variable but mostly around the previously estimated genome sizes [39, 43]. Based on results from the *A. thaliana* investigation, the mean calculation among all single copy BUSCOs should be the best approach. The prediction of slightly less than 600 Mbp is probably an underestimation, but still the highest reliable estimate. When assuming centromere sizes of only 2-3 Mbp per chromosome, this number could be in a plausible range. However, a previous investigation of the repeat content indicates a larger genome size due to a high number of repeats which are not represented in the assembly [58].

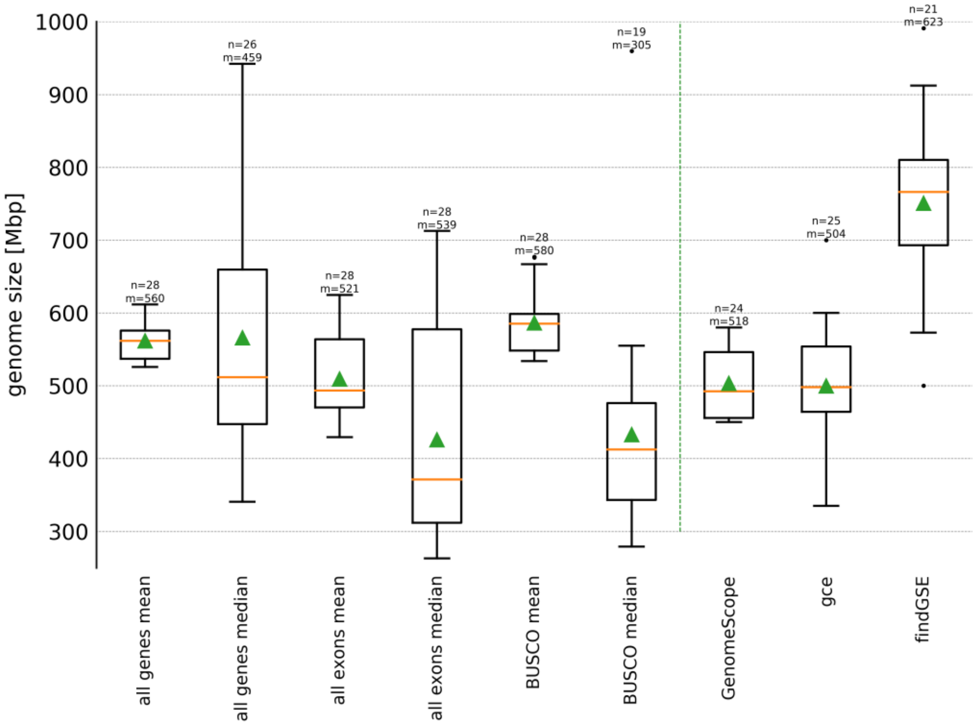

Fig. 3: Comparison of *Beta vulgaris* genome size estimations.

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).

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 heterozigosity. The predictions of MGSE are generally in the same range as the predictions generated by GenomeScope, gce, and findGSE (AdditionalFile5, AdditionalFile6, AdditionalFile7, AdditionalFile8, and AdditionalFile9). With an average prediction of 290 Mbp as genome size of *B. distachyon*, the MGSE prediction is slightly exceeding the assembly size. GenomeScope and gce predict genome sizes below the assembly size, while the prediction of 303 Mbp by findGSE is more reasonable. The *Z. mays* genome size is underestimated by all four tools. However, MGSE outperforms GenomeScope and gce on the analyzed data set. 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. The highest MGSE predictions are in the range of the expected genome size. MGSE failed for *V. vinifera* by predicting only 50 Mbp. The high heterozigosity of this species could contribute to this by causing lower mapping rates outside of important protein encoding genes i.e. BUSCO genes.

Considerations about performance and outlook

MGSE performs best on a high contiguity assembly and requires a (short) read mapping 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 is the apparent coverage of this contig 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 [23, 24, 59, 60]. Future genome projects are likely to generate high continuity assemblies and short 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 expects 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 reads is usually small, such reads are frequently observed due to the high sensitivity of next generation sequencing [31, 61–64].

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 GMSE 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 [65, 66]. 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 [67, 68]. Both, extremely GC-rich and AT-rich fragments, respectively, are underrepresented in the sequencing output mainly due to biases introduced by PCR [69, 70]. Sophisticated methods were developed to correct coverage values based on the GC content of the underlying sequence [70–72]. 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 [56]. 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 are necessary to develop a correction factor for this GC bias of reads.

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.

The major issue when developing tools for the genome size prediction is the absence of a gold standard. Since as of yet there is no completely sequenced plant genome, benchmarking with real data cannot be perfect. As a result, how various estimation approaches will compare to the first completely sequenced and assembled genome remains speculative. 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 are freely available on github: https://github.com/bpucker/MGSE (https://doi.org/10.5281/zenodo.2636733). Underlying data sets are publicly available at the NCBI and SRA, respectively.

Acknowledgements

Members of Genetics and Genomics of Plants 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.

References

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

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

3. Bennett MD, Smith JB. Nuclear DNA Amounts in Angiosperms. Philos Trans Biol Sci. 1991;334:309–45.

4. Francis DM, Hulbert SH, Michelmore RW. Genome size and complexity of the obligate fungal pathogen, *Bremia lactucae*. Exp Mycol. 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 Report. 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. Nat Rev Genet. 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. ArXiv160905626 Cs. 2016. http://arxiv.org/abs/1609.05626. Accessed 10 Feb 2019.

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. ArXiv13082012 Q-Bio. 2013. http://arxiv.org/abs/1308.2012. Accessed 10 Feb 2019.

16. 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.

17. 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. Proc Natl Acad Sci. 2011;108:1513–8.

18. 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.

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

20. 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. Proc Natl Acad Sci. 2016;113:E4052–60.

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

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

23. 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. Nat Commun. 2018;9:541.

24. Pucker B, Holtgraewe 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. bioRxiv 407627; doi: https://doi.org/10.1101/407627.

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

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

27. 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. doi:10.1038/sdata.2014.45.

28. 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.

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

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

31. 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.

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

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

34. 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. Mol Plant. 2015;8:1651–64.

35. 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.

36. 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.

37. Wynn E, Christensen A. Do Plant Mitochondria Even Need Base Excision Repair? bioRxiv. 2018;:427500.

38. 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. Plant Cell. 2018;30:1954–70.

39. 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.

40. 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.). Front Plant Sci. 2016;7. doi:10.3389/fpls.2016.01662.

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

42. 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.

43. Holtgräwe D, Rosleff Sörensen T, Parol-Kryger R, Pucker B, Kleinbölting N, Viehöver P, et al. Low coverage re-sequencing in sugar beet for anchoring assembly sequences to genomic positions. 2017. https://jbrowse.cebitec.uni-bielefeld.de/RefBeet1.5/.

44. Stadermann KB, Weisshaar B, Holtgräwe D. SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. BMC Bioinformatics. 2015;16. doi:10.1186/s12859-015-0726-6.

45. 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.

46. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010;463:763–8.

47. 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.

48. 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 Genet. 2018;14:e1007807.

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

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

51. 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 Biol. 2013;14:R33.

52. 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.

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

54. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013. http://arxiv.org/abs/1303.3997. Accessed 16 Oct 2018.

55. Picard Tools - By Broad Institute. https://broadinstitute.github.io/picard/. Accessed 10 Feb 2019.

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

57. 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.

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

59. 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;:gr.213652.116.

60. Saint-Oyant LH, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nat Plants. 2018;4:473.

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

62. 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 Biol. 2014;12:87.

63. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, et al. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. PLoS Pathog. 2014;10. doi:10.1371/journal.ppat.1004437.

64. Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H. PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. Bioinformatics. 2017;33:3283–5.

65. 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.

66. 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. 2013;111:314–20.

67. 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.

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

69. 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 Biol. 2011;12:R18.

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

71. 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.

72. 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.

Supplements:

AdditionalFile1: Sequencing data set overview.

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

AdditionalFile3: *A. thaliana* genome size prediction based on Ath-Nd1_v1.

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

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

AdditionalFile6: Genome size estimation of *Brachypodium distachyon*.

AdditionalFile7: Genome size estimation of *Zea mays*.

AdditionalFile8: Genome size estimation of *Solanum lycopersicum*.

AdditionalFile9: Genome size estimation of *Vitis vinifera*.