Mabs, a suite of tools for gene-informed genome assembly

Author information

Mikhail I. Schelkunov[1,2]

[1]Institute for Information Transmission Problems, Moscow, Russia

[2]Skolkovo Institute of Science and Technology, Moscow, Russia

e-mail: shelkmike@gmail.com

Abstract

Motivation: Despite constantly improving methods of genome sequencing, the task of error-free eukaryotic genome assembly is not yet solved. Among other kinds of problems of eukaryotic genome assembly are so-called "haplotypic duplications", which may manifest themselves as cases when alleles are mistakenly assembled as paralogs. Haplotypic duplications are dangerous as they create illusions of gene family expansions and, thus, may bring a scientist to wrong conclusions about genome evolution and functioning.

Results: Here, I present Mabs, a suite of tools that serve as parameter optimizers of popular genome assemblers Hifiasm and Flye. By optimizing parameters of Hifiasm and Flye, Mabs tries to create assemblies where genes are assembled as accurately as possible. Tests on 6 eukaryotic genomes showed that in 6 out of 6 cases Mabs created assemblies with more accurately assembled genes than Hifiasm and Flye did when they were run with default parameters. When assemblies of Mabs, Hifiasm and Flye were postprocessed by a popular tool for haplotypic duplications removal Purge_dups, genes were better assembled by Mabs in 5 out of 6 cases.

Availability and implementation: Mabs has been written in Python and is available at https://github.com/shelkmike/Mabs

1 Introduction

In recent years, sequencing technologies have improved significantly. Reads of Oxford Nanopore Technologies have become longer and more accurate (Wang *et al.*, 2021), and so did HiFi reads of PacBio (Pacific Biosciences, 2019, 2021). Despite this progress, genome assemblies still suffer from a number of problems, among the major of which are:

1. Fragmentation owing to long repeats with similar copies (Dida and Yi, 2021; Nurk *et al.*, 2022).
2. Contamination (Steinegger and Salzberg, 2020; Cornet and Baurain, 2022).
3. Haplotypic duplications (Guan *et al.*, 2020; Ko *et al.*, 2022).

The latter problem is a case when, during an assembly of a diploid or a polyploid genome, a genome assembler mistakes corresponding regions of two homologous chromosomes with regions that originated from segmental duplications. For example, two alleles of the same gene may be mistaken for paralogs. Haplotypic duplications are dangerous as they may lead to incorrect scientific conclusions about the gene content of a genome. When two alleles are assembled separately, as paralogs, an illusion of a gene duplication is created. In highly

heterozygous genomes haplotypic duplications are so massive that they may create such false duplicates for thousands of genes (Guan *et al.*, 2020; Ko *et al.*, 2022).

One way to deal with haplotypic duplications is to minimize them during the process of genome assembly. For example, authors of a genome assembler Hifiasm endowed it with a special algorithm that tells apart corresponding regions of homologous chromosomes and segmental duplications (Cheng *et al.*, 2021). An alternative method is to try to remove haplotypic duplications after the assembly. This method is implemented in specialized programs like Purge_dups (Guan *et al.*, 2020), Purge_haplotigs (Roach *et al.*, 2018) and HapSolo (Solares *et al.*, 2021). However, there are yet no methods or combinations of methods that are 100% effective in removal of haplotypic duplications.

To deal with the problem of haplotypic duplications I created a suite of tools called "Mabs" that I describe in this article. The main two components of Mabs are Mabs-hifiasm and Mabs-flye that serve as parameter optimizers of popular genome assemblers Hifiasm (Cheng *et al.*, 2021) and Flye (Kolmogorov *et al.*, 2019). Mabs-hifiasm is intended for assembly using PacBio HiFi (also known as PacBio CCS) reads, while Mabs-flye is intended for assembly using reads of more error-prone technologies, namely Oxford Nanopore Technologies and PacBio CLR. By optimizing parameters of Hifiasm or Flye, Mabs reduces the number of haplotypic duplications.

2 Materials and Methods

2.1 Overview of the algorithm

For a detailed description of the algorithm, see Supplementary Text S1.

Mabs tries to find values of parameters of a genome assembler that maximize the number of accurately assembled BUSCO genes. BUSCO is a program that is supplied with a number of taxon-specific datasets that contain orthogroups whose genes are present and single-copy in at least 90% of genomes from a given taxon (Simão *et al.*, 2015). Thus, it is expected that most BUSCO genes in any newly studied genome will also be present and single-copy. Mabs discriminates found BUSCO genes into three categories:

1. Belonging to single-copy orthogroups.
2. Belonging to false multicopy orthogroups. These genes are haplotypic duplications.
3. Belonging to true multicopy orthogroups. These genes are paralogs.

The original algorithm of BUSCO does not differentiate between false multicopy and true multicopy orthogroups, combining all orthogroups into a single category "D" (short for "duplicated"). However, to properly evaluate the quality of a genome assembly, it is preferential to be able to tell apart false multicopy from true multicopy orthogroups. To do this, Mabs uses the following procedure (see also Fig. 1A):

1. Calculate the median coverage in exons of BUSCO genes from single-copy orthogroups. I denote it Cov(S).
2. Calculate the median coverage in exons of BUSCO genes from multicopy orthogroups. The coverage is calculated separately for each BUSCO orthogroup.
3. If the median coverage in exons of genes of a multicopy BUSCO orthogroup is closer to Cov(S) than to (Cov(S) / 2), then genes of this orthogroup are considered true multicopy. Otherwise, they are considered false multicopy. This method uses the fact that the

coverage of alleles not collapsed during the assembly should be two times lower than the coverage of collapsed alleles.

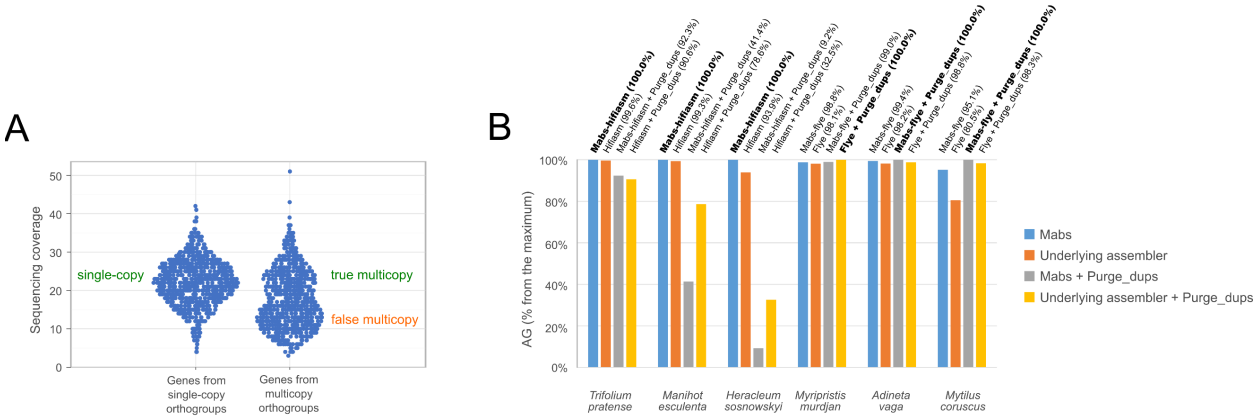
Mabs uses a metric that I call AG (short for "Accurately assembled Genes"), which is calculated as the sum of the number of BUSCO genes in single-copy orthogroups and the number of BUSCO genes in true multicopy orthogroups. Basically, Mabs works as a parameter optimizer of a genome assembler, trying to find such values of parameters that maximize AG. An underlying assumption of Mabs is that optimization of the assembly quality of BUSCO genes leads to optimization of assembly quality of other protein-coding genes too. Thus, by maximizing AG, Mabs tries to create a genome assembly where protein-coding genes are assembled as well as possible.

In principle, this strategy of "gene-informed parameter optimization" may be applied to any genome assembler that has some parameters which affect its algorithm. I have chosen Hifiasm and Flye because in multiple benchmarks they were shown to be the best or among the best assemblers for accurate (PacBio HiFi) and error-prone (Oxford Nanopore Technologies and PacBio CLR) reads respectively (Jung *et al.*, 2020; Murigneux *et al.*, 2020; Dida and Yi, 2021; Gavrielatos *et al.*, 2021; Guiglielmoni *et al.*, 2021; Schneider *et al.*, 2021; Wick and Holt, 2021; Rabanal *et al.*, 2022; Xie *et al.*, 2022; Zhang *et al.*, 2022). Since Hifiasm and Flye were favorably compared with other genome assemblers many times, in this work I compare Mabs only with Hifiasm and Flye and do not make comparisons with other assemblers.

2.2 Genomes

To test Mabs, I performed a literature analysis and compiled a set of 5 species for which authors reported a high number of haplotypic duplications: *Trifolium pratense* (Bickhart *et al.*, 2022), *Manihot esculenta* (Qi *et al.*, 2022), *Myripristis murdjan* (Guan *et al.*, 2020; Rhie *et al.*, 2021), *Adineta vaga* (Guiglielmoni *et al.*, 2021), *Mytilus coruscus* (Li *et al.*, 2020; Sun *et al.*, 2021). In addition, I used *Heracleum sosnowskyi*, which I studied personally and which also had a lot of haplotypic duplications (preprint to be uploaded to BioRxiv in 2022). For detailed information about the genomes and their sequencing reads, see Supplementary Tables S1 and S2. The genomes of *Trifolium pratense*, *Manihot esculenta* and *Heracleum sosnowskyi* were sequenced using the PacBio HiFi technology and were thus used to compare Mabs-hifiasm and Hifiasm. The other three genomes were sequenced using error-prone technologies and were thus used to compare Mabs-flye and Flye.

Fig. 1



(**A**) Sinaplots (Sidiropoulos *et al.*, 2018) with coverage of BUSCO genes in the Hifiasm assembly of the genome of *Heracleum sosnowskyi*. Every point is a gene. For multicopy orthogroups, separation into true multicopy (high coverage) and false multicopy (low coverage) can be seen. (**B**) Values of AG for assemblies of the six genomes. AG is indicated per cent with respect to the maximum value achieved for this genome. The best assemblies are marked in bold.

3 Results

A detailed comparison of assemblies made by Mabs and its underlying assemblers Hifiasm and Flye is provided in Supplementary Table S3.

If AG is used as the metric of gene assembly quality, for all 6 genomes Mabs created better assemblies than Hifiasm and Flye (Fig. 1B). Larger AG implies a higher quality of assembly of protein-coding genes. The larger amount of haplotypic duplications in assemblies made by Hifiasm and Flye can also be seen as larger values of BUSCO's "D" (the number of multicopy orthogroups), see Supplementary Table S3. In several most prominent cases, the larger amount of haplotypic duplications made by the underlying assemblers compared with Mabs can be seen through diagrams of gene coverage, see Supplementary Figs. S1 and S2. In terms of N50, assemblies made by Mabs were better than assemblies made by Hifiasm and Flye for 5 of 6 genomes.

When Purge_dups was applied after the assembly, assemblies made by Mabs were better than assemblies of Hifiasm and Flye for 5 of 6 genomes (Supplementary Table S3). Overall, Purge_dups was detrimental (it decreased AG) for all assemblies from HiFi reads, but was beneficial for all assemblies from error-prone reads.

Despite the higher assembly quality, for all 6 genomes Mabs was slower than the underlying assemblers, with the assembly time of Mabs being 1.6-3.5 times larger than theirs. RAM consumption was approximately the same as in the underlying assemblers (Supplementary Table S3).

4 Conclusions

In this article, I demonstrated the utility of a tool Mabs that optimizes parameters of genome assemblers Hifiasm or Flye. The idea of "gene-informed parameter optimization" used by Mabs may potentially be implemented with other genome, metagenome or transcriptome assemblers as well.

5 Acknowledgements

The author is thankful to Maria D. Logacheva, Aleksey A. Penin, Artem S. Kasianov and Maksim S. Makarenko for their valuable comments.

The development of Mabs was funded by the budgetary subsidy with the code FFNU-2022-0037 to the Institute for Information Transmission Problems of the Russian Academy of Sciences. The genome assembly for *Heracleum sosnowskyi* was funded by the research grant no. 21-74-20145 of the Russian Science Foundation.

References

Bickhart,D.M. *et al.* (2022) Chromosome-scale assembly of the highly heterozygous genome of red clover (Trifolium pratense L.), an allogamous forage crop species. *Gigabyte*, **2022**, 1–13.

Cheng,H. *et al.* (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, **18**, 170–175.

Cornet,L. and Baurain,D. (2022) Contamination detection in genomic data: more is not enough. *Genome Biol*, **23**, 60.

Dida,F. and Yi,G. (2021) Empirical evaluation of methods for de novo genome assembly. *PeerJ Comput Sci*, **7**, e636.

Gavrielatos,M. *et al.* (2021) Benchmarking of next and third generation sequencing technologies and their associated algorithms for *de novo* genome assembly. *Mol Med Rep*, **23**, 251.

Guan,D. *et al.* (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896–2898.

Guiglielmoni,N. *et al.* (2021) Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics*, **22**, 303.

Jung,H. *et al.* (2020) Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. *J. Agric. Food Chem.*, **68**, 7670–7677.

Ko,B.J. *et al.* (2022) Widespread false gene gains caused by duplication errors in genome assemblies. *Genome Biol*, **23**, 205.

Kolmogorov,M. *et al.* (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, **37**, 540–546.

Li,R. *et al.* (2020) The Whole-Genome Sequencing and Hybrid Assembly of Mytilus coruscus. *Front. Genet.*, **11**, 440.

Murigneux,V. *et al.* (2020) Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*, **9**, giaa146.

Nurk,S. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.

Pacific Biosciences (2021) Pacific Biosciences Launches New HiFi Sequencing Workflow to Further Improve HiFi's Industry Leading Accuracy (https://www.pacb.com/press_releases/pacific-biosciences-launches-new-hifi-sequencing-workflow-to-further-improve-hifis-industry-leading-accuracy/).

Pacific Biosciences (2019) Sequel II System 2.0 Chemistry and v8.0 Software Release (https://www.pacb.com/technology/hifi-sequencing/sequel-system/previous-system-releases/).

Qi,W. *et al.* (2022) The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *GigaScience*, **11**, giac028.

Rabanal,F.A. *et al.* (2022) Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes Genomics.

Rhie,A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.

Roach,M.J. *et al.* (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 460.

Schneider,C. *et al.* (2021) Two high-quality *de novo* genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *GigaScience*, **10**, giab035.

Sidiropoulos,N. *et al.* (2018) SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations Over Multiple Classes. *Journal of Computational and Graphical Statistics*, **27**, 673–676.

Simão,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Solares,E.A. *et al.* (2021) HapSolo: an optimization approach for removing secondary haplotigs during diploid genome assembly and scaffolding. *BMC Bioinformatics*, **22**, 9.

Steinegger,M. and Salzberg,S.L. (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol*, **21**, 115.

Sun,J. *et al.* (2021) Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Phil. Trans. R. Soc. B*, **376**, 20200160.

Wang,Yunhao *et al.* (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*, **39**, 1348–1365.

Wick,R.R. and Holt,K.E. (2021) Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res*, **8**, 2138.

Xie,H. *et al.* (2022) *De novo* assembly of human genome at single-cell levels. *Nucleic Acids Research*, **50**, 7479–7492.

Zhang,X. *et al.* (2022) Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. *Briefings in Bioinformatics*, **23**, bbac146.