Benchmarking of Hi-C tools for scaffolding de novo genome assemblies

Lia Obinu [1]*, Urmi Trivedi [2], Andrea Porceddu [1]

[1] Department of Agricultural Sciences, University of Sassari, Viale Italia 39A, 07100, Sassari, Sardinia, Italy

[2] Edinburgh Genomics, Ashworth Laboratories, Charlotte Auerbach Road, The King's Buildings, The University of Edinburgh, EH9 3FL, Edinburgh, Scotland, United Kingdom

*Corresponding Author: lobinu@uniss.it

Abstract

The implementation of Hi-C reads in the *de novo* genome assembly allows to order large regions of the genome in scaffolds, obtaining chromosome-level assemblies. Several bioinformatics tools have been developed for genome scaffolding with HiC, and all have pros and cons which need to be carefully evaluated before adoption.

Here we considered QUAST and BUSCO scores to benchmark the performance of three scaffolders, 3d-dna, SALSA2, and YaHS, on a *de novo* assembly of *Arabidopsis thaliana.*

In our analysis, YaHS resulted to be the best performing bioinformatic tool for scaffolding of *de novo* genome assembly.

Introduction

Third-generation sequencing technologies (TGS), e.g., Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), produce long reads which allow to obtain high-quality genome sequences. Compared to next generation sequencing technologies, TGS give better resolution and contiguity, allowing to partially resolve the problems of repeats and duplications, which are crucial to resolve repetitive rich genomic regions such as the centromeres and telomeres, especially in plant genomes[1].

Although long read assembly on itself can yield good length contigs, this can still be far from assembling to a chromosome scale. Strategically, larger genome projects like Vertebrate Genome Project (VGP - https://vertebrategenomesproject.org/), The Darwin Tree of Life (https://www.darwintreeoflife.org/) combine multiple data types like optical mapping[2] and Hi-C[3] to construct reliably chromosome-scale assemblies[4]. A further step ahead in genome assembly has been accomplished by implementing Hi-C reads in *de novo* genome assembly. Hi-C technology is a method which combines proximity-based ligation with massively parallel sequencing allowing the unbiased identification of chromatin interactions across an entire genome. Hi-C reads have been developed to study the three-dimensional organisation of the genome, by measuring the frequency of contact between pairs of loci[5]. Currently, Hi-C reads are largely used to improve genome assemblies by scaffolding together large regions of the genome[6], obtaining chromosome-level assemblies[7].

Previous study about Hi-C based scaffolders benchmarking was conducted on partitioned reference genomes and *de novo* assemblies[8]. The scaffolders' performance evaluation was based on the mapping of assemblies against the known reference genome using Mummer 4[9], and on a set of accuracy metrics calculated using the python package Edison (https://github.com/Noble-Lab/edison).

In this study, we benchmarked three Hi-C scaffolders (3d-dna[10] , SALSA2[11], and YaHS[12]) starting from a *de novo* assembly of *A. thaliana,* the model species for plant genomics[13]. Our benchmarking is based on QUAST[14] and BUSCO[15] metrics.

3d-dna was chosen for being the most applied Hi-C scaffolder still under active development[8]. SALSA2 is the new version of SALSA[16], while YaHS is the most recently released scaffolder. Indeed, the last two scaffolders were never benchmarked with other scaffolders before.

Material and methods

Arabidopsis thaliana genome and raw data

A. thaliana genome was sequenced for the first time in 2000, using the method of the minimum tiling path of BACs sequenced with Sanger's technology[17]. The first reference sequence was obtained for the Columbia genotype and it is approximately 135 Mb in length, organised in 5 chromosomes (2n = 10).

In this study, we used the publicly available raw data from the BioProject PRJCA005809[18] to obtain a *de novo* assembly and evaluate the performance of the three scaffolders.

Table 1 summarises the characteristics of the raw reads used.

Table 1. Overview of the characteristics of the raw reads from the BioProject PRJCA005809[18] used to assemble the draft assembly.

Reads	Mean read length	Mean read quality	Number of reads	Total bases	Coverage
ONT	18,541.30	11.1	3,064,191.00	56,814,196,989.00	420.84507
PacBio HiFi	15,094.40	31.8	1,517,433.00	22,904,700,074.00	169.66475
Hi-C	150	>30	70478750	10571812500	78.309722

De novo assembly

ONT reads were trimmed with NanoFilt[19] using the parameter -l 500, which only retained the reads with minimum length of 500 bp. The *de novo* assembly was obtained assembling ONT reads using Flye[20] in --nano-raw mode with default parameters. To subsequently perform the polishing, the PacBio HiFi reads were mapped to the draft assembly using minimap2[21] in map-hifi mode, and then the polishing was undertaken with Racon[22]. Hereafter, we will refer to this assembly as the "draft assembly".

Hi-C scaffolding

All Hi-C scaffolders, namely 3d-dna, YaHS and SALSA2 were run to scaffold the draft assembly using default parameters.

3d-dna

3d-dna has been developed to produce chromosome-level scaffolds using Hi-C linking information. To be used with 3d-dna, Hi-C reads must be pre-processed, i.e. filtered and aligned to the draft assembly, using Juicer[23], which produces a list of Hi-C contacts in txt format. The inputs required by the software are only the output of Juicer and the fasta file of the draft assembly. The Hi-C are used to correct the errors in the draft assembly, and then the resulting sequences are anchored, ordered and oriented thanks to an algorithm based on the contact frequency between pairs of sequences as an indicator of their proximity in the genome. Finally, contigs and scaffolds that correspond to overlapping regions of the genome are merged together[10].

SALSA2

SALSA2 is an open-source software that combines Hi-C linkage information with draft genome assemblies – i.e. contig level assemblies generated from PacBio or ONT long reads – to resolve contig orientations.

The Hi-C reads must be previously filtered and aligned to the draft assembly following the Arima Genomics pipeline (https://github.com/ArimaGenomics/mapping_pipeline). SALSA2 requires as inputs the fasta and fasta.fai files of the draft assembly and the Arima Genomics pipeline output in bed format. A scaffold graph is constructed scoring edges according to a "best buddy" scheme. The scaffolds are then constructed according to a greedy weighted maximum matching. Afterwards, SALSA2 performs an iterative step of mis-join detection and correction, which stops naturally when accurate Hi-C links are exhausted[11].

YaHS

YaHS is the newest scaffolder that has been developed and, compared to the previous ones, it proposes a new method for building the contact matrix. The Hi-C reads must be quality filtered and mapped to the draft assembly using the Arima Genomics pipeline (https://github.com/ArimaGenomics/mapping_pipeline) before proceeding with the scaffolding. To run, YaHS needs as inputs the Arima Genomics pipeline output in bam or bed format and the fasta file of the draft assembly and its index (fasta.fai file). The software proceeds building a contact matrix, constructing and pruning a scaffold graph, and giving the output scaffolds[12].

The full commands used are shown in supplementary materials (S1).

Evaluation criteria

The three scaffolded assemblies were evaluated according to QUAST[14] and BUSCO v5.4.2[15] metrics.

QUAST is a tool used to compare *de novo* assemblies against reference sequences, which computes relevant quality metrics. The evaluation of the scaffolded assemblies was based on the QUAST metrics of genome fraction, number of contigs, N50, N90, L50, and L90.

BUSCO evaluates genome assembly quality in terms of expected gene content. The scaffolded assemblies were compared with the database for Brassicales (brassicales_odb10) and evaluated considering all the metrics provided by BUSCO (Complete BUSCOs, Complete and single-copy BUSCOs, Complete and duplicated BUSCOs, Fragmented BUSCOs, and Missing BUSCOs).

Cluster characteristics

The programmes in benchmarking were executed on a linux machine containing two 2.1 GHz, 18-core Intel Xeon E5-2695 (Broadwell) series processors where each of the cores in these processors support 2 hyperthreads enabled by default with 256 GB of memory.

Results and discussion

JuiceBox Assembly Tools (JBAT)[24] allows to visualise the contact map for the assemblies scaffolded using 3d-dna and YaHS (Figure 1). SALSA2 doesn't provide a valid method that allows to visualise the contact map in JBAT.

We observed that the major scaffolds corresponding to the chromosomes were correctly identified by 3d-dna, which identified 5 major scaffolds (Figure 1A). The assembly scaffolded with YaHS identified 4 major scaffolds, because two of the chromosome-length scaffolds were joined (Figure 1B). However, the error could be easily detected and corrected with JBAT, by manually splitting the two chromosome length scaffolds (Figure 1C), as explained in the Aiden Lab's Genome Assembly Cookbook (https://www.dnazoo.org/methods). The JBAT corrected assembly is here referred as YaHS_corrected.

Figure 2 shows the QUAST metrics of the scaffolded assemblies compared to the draft assembly. As regards YaHS, we will take into consideration the manually corrected assembly (YaHS_corrected), to compare the performance of the software with the others.

All the scaffolders were able to align more than 99% of bases to the reference genome of *A. thaliana* (TAIR10).

SALSA2 and YaHS were able to reduce the number of contigs of the draft assembly, while 3d-dna increased it. The best result was achieved by YaHS, which reduced the number of contigs from 42 to 33.

All the scaffolders increased the N50 compared to the draft assembly. As regards the N90, the assembly scaffolded with SALSA2 had exactly the same result as the draft assembly (9.47 Mb). YaHS had both the highest N50 (24.26 Mb) and N90 (19.11 Mb).

YaHS also achieved the lowest L50 (3) and L90 (5).

The full QUAST report is shown in supplementary materials (S2).

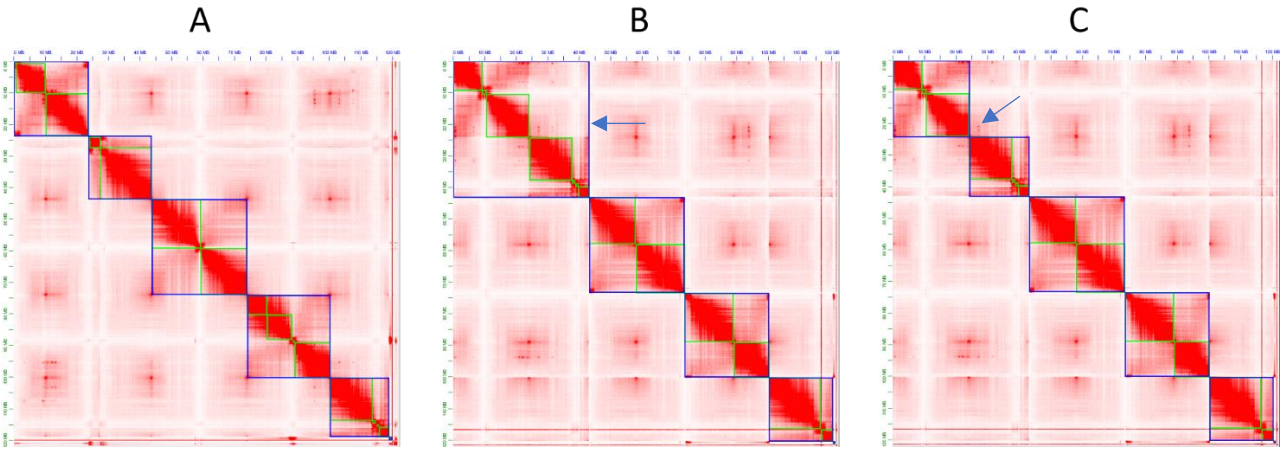

Figure 1. Contact maps obtained with JBAT. A shows the contact map of the assembly scaffolded with 3d-dna. B shows the contact map obtained scaffolding the assembly with YaHS. C shows the contact map obtained scaffolding the assembly with YaHS after the manual correction of the boundary between the first and the second chromosome-length scaffolds. The arrow in B and C highlights the chromosome boundary that has been manually corrected.

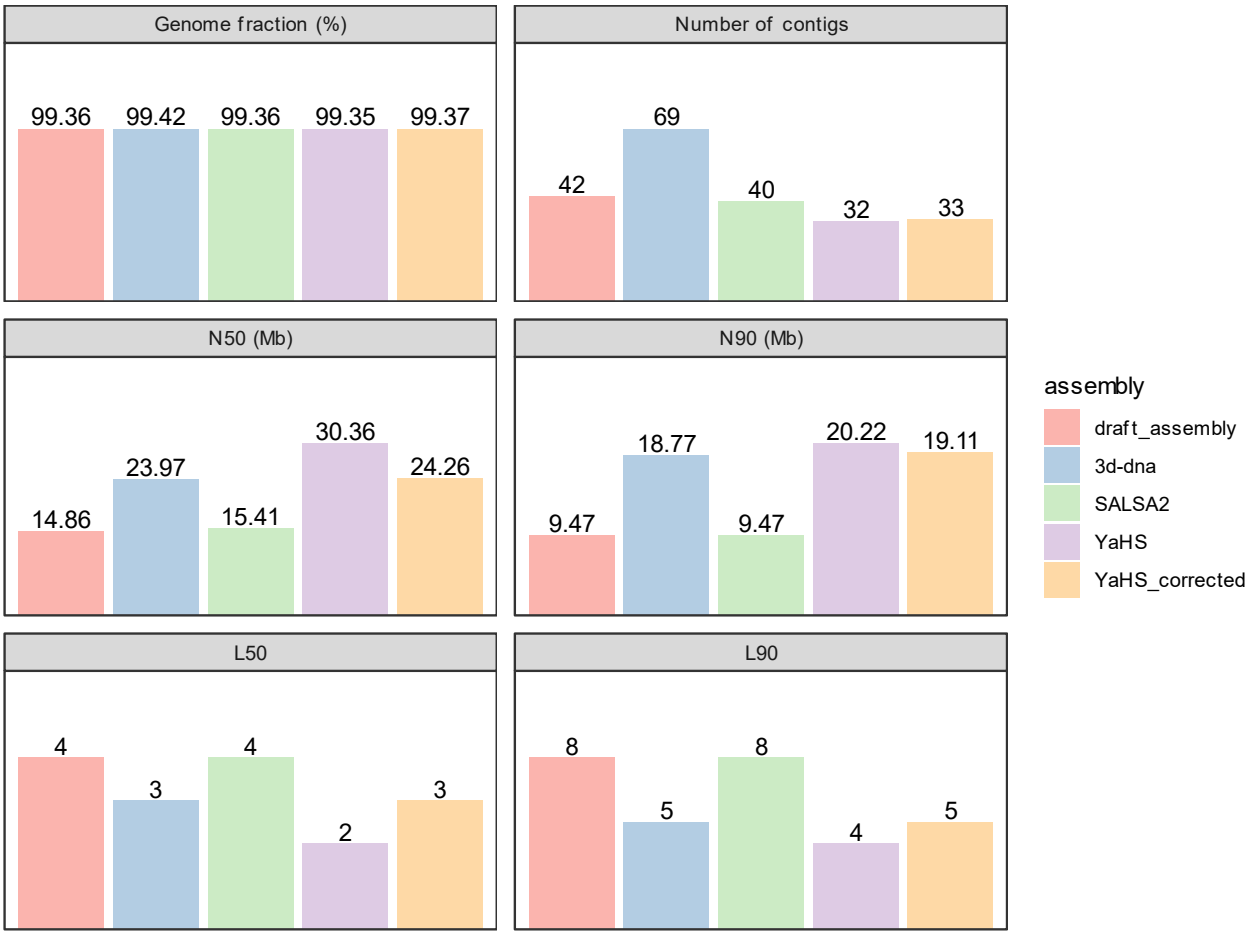

Figure 2. Comparison of assemblies according to QUAST metrics.

Table 2 shows the comparison between the draft assembly and the scaffolded assemblies as regard the BUSCO metrics.

YaHS was the only scaffolder able to increase the number of complete BUSCOs and complete and single-copy BUSCOs. 3d-dna and SALSA2 provided results equal or slightly worse compared to the draft assembly.

YaHS performed better than 3d-dna and SALSA2 also as regards complete and deduplicated BUSCOs and missing BUSCOs, giving the lowest results in both cases. 3d-dna couldn't improve the assembly in respect of the same metrics, while SALSA2 slightly reduced the number of complete and duplicated BUSCOs.

Concerning the fragmentation of the assembly (fragmented BUSCOs), all the scaffolders were able to keep the initial fragmentation of the draft assembly.

Table 2. Comparison of assemblies according to BUSCO metrics.

	draft_assembly	**3d-dna**	**Salsa2**	**YaHS**	**YaHS_corrected**
Complete BUSCOs	4559	4558	4558	4560	4560
Complete and single-copy BUSCOs	4495	4495	4494	4500	4498
Complete and duplicated BUSCOs	64	63	64	60	62
Fragmented BUSCOs	2	2	2	2	2
Missing BUSCOs	35	36	36	34	34
Total BUSCO groups searched	4596	4596	4596	4596	4596

Overall, YaHS performed better than both SALSA2 and 3d-dna as regards QUAST and BUSCO scores. 3d-dna achieved better results compared to SALSA2, providing better metrics for N50, N90, L50, L90, complete and single copy BUSCOs, and complete and duplicated BUSCOs. SALSA2 was better than 3d-dna only as regards the number of contigs.

YaHS was the easiest and quicker software to install and use, taking into account also the Hi-C pre-processing. In addition, the developers provide good documentation to implement the software in the analysis. 3d-dna also has good documentation, but we found that it is more difficult to install and run. As regard SALSA2, the main flaws are the lack of a detailed and rich documentation, and the impossibility of visualising a contact map in JBAT with the methods provided.

Conclusions

Of the three scaffolders that were benchmarked in this study, YaHS was the best performing for the parameters considered. Therefore, we conclude that it is the most appropriate scaffolding tool for *de novo* assemblies to date.

Acknowledgments

We would like to thank Dr. Alex Mackintosh of The University of Edinburgh for his useful advice and inputs on the topic.

Fundings

This study includes part of a PhD project carried out by Lia Obinu at the PhD School of Agricultural Sciences of the University of Sassari, funded by the University of Sassari. Lia Obinu was also supported by ERASMUS for Traineeship program awarded to the University of Sassari.

References

1. Jung, H., Winefield, C., Bombarely, A., Prentis, P. & Waterhouse, P. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends Pla*1. Jung, H., Winefield, C., Bombarely, A., Prentis, P. & Waterhouse, P. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends Plant Sci.* **24**, 700–724 (2019).

2. Zhou, S., Herschleb, J. & Schwartz, D. C. Chapter 9 A Single Molecule System for Whole Genome Analysis. *Perspect. Bioanal.* **2**, 265–300 (2007).

3. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).

4. Guan, D. *et al.* Efficient iterative Hi-C scaffolder based on N-best neighbors. *BMC Bioinformatics* **22**, 1–16 (2021).

5. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. **33292**, 289–294 (2009).

6. Ghurye, J. & Pop, M. Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Comput. Biol.* **15**, 1–20 (2019).

7. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 1–10 (2014).

8. Sur, A., Noble, W. S. & Myler, P. J. A benchmark of Hi-C scaffolders using reference genomes and de novo assemblies. *bioRxiv* 2022.04.20.488415 (2022).

9. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, 1–14 (2018).

10. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (80-.).* **356**, 92–95 (2017).

11. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, 1–18 (2019).

12. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *bioRxiv* 2022.06.09.495093 (2022).

13. Koornneef, M. & Meinke, D. The development of Arabidopsis as a model plant. *Plant J.* **61**, 909–921 (2010).

14. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

15. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

16. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 1–11 (2017).

17. AGI *et al.* Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* vol. 408 796–815 (2000).

18. Wang, B. *et al.* High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. *Genomics. Proteomics Bioinformatics* (2021) doi:10.1016/j.gpb.2021.08.003.

19. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

20. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

21. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

22. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

23. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).

24. Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv* 254797 (2018).