Bioinformatics, 2025, **41(5)**, btaf175
https://doi.org/10.1093/bioinformatics/btaf175
Advance Access Publication Date: 16 April 2025
Original Paper

OXFORD

Genome analysis

Colora: a Snakemake workflow for complete chromosome-scale *de novo* genome assembly

Lia Obinu[1,2,*], Timothy Booth[2], Heleen De Weerd[2], Urmi Trivedi[2], Andrea Porceddu[1]

[1]Department of Agricultural Sciences, University of Sassari, Viale Italia 39/a, Sassari, Sardinia, 07100, Italy
[2]Edinburgh Genomics, The University of Edinburgh, Ashworth Laboratories, The King's Buildings, Charlotte Auerbach Rd, Edinburgh, Scotland, EH9 3FL, United Kingdom

*Corresponding author. Department of Agriculture, University of Sassari, Viale Italia 39/a, Sassari, Sardinia, 07100, Italy. E-mail: lobinu@uniss.it.

Associate Editor: Peter Robinson

Abstract

Motivation: *De novo* assembly creates reference genomes that underpin many modern biodiversity and conservation studies. Large numbers of new genomes are being assembled by labs around the world. To avoid duplication of efforts and variable data quality, we desire a best-practice assembly process, implemented as an automated portable workflow.

Results: Here, we present Colora, a Snakemake workflow that produces chromosome-scale *de novo* primary or phased genome assemblies complete with organelles using Pacific Biosciences HiFi, Hi-C, and optionally Oxford Nanopore Technologies reads as input. Colora is a user-friendly, versatile, and reproducible pipeline that is ready to use by researchers looking for an automated way to obtain high-quality *de novo* genome assemblies.

Availability and implementation: The source code of Colora is available on GitHub (https://github.com/LiaOb21/colora) and has been deposited in Zenodo under DOI https://doi.org/10.5281/zenodo.13321576. Colora is also available at the Snakemake Workflow Catalog (https://snakemake.github.io/snakemake-workflow-catalog/? usage=LiaOb21%2Fcolora).

1 Introduction

Third-generation sequencing technologies, such as the platforms from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can be used to obtain high-quality *de novo* genome assemblies for model and non-model species (Jayakumar and Sakakibara 2019, Jung *et al.* 2019). However, long-read assemblies alone are unlikely to create chromosome-scale assemblies, and, therefore, Hi-C reads are widely used to perform assembly scaffolding (Guan *et al.* 2021).

Currently, *de novo* genome assemblies are the basis of cutting-edge biodiversity and conservation studies. However, the bioinformatics procedures used to obtain such results are often challenging in terms of time, study, and human and computational resources. This is particularly true for small labs, which do not have access to the same resources that are available to big institutions. Therefore, automating workflows by combining up-to-date techniques and tools in an easily implementable, modifiable, and portable way is fundamental.

Snakemake is a popular Python-based workflow engine that allows the automation of workflows for single machines, such as personal laptops, and High-Performance Computer (HPC) clusters (Köster and Rahmann 2012). It allows the creation of separate environments for each bioinformatics tool through Conda (Anaconda Software Distribution 2024, https://anaconda.com), Docker (Merkel 2014) or Singularity (Kurtzer *et al.* 2017), making the installation of the whole pipeline automatic and thus easy for the user and ensuring reproducibility.

Here, we present Colora, a Snakemake workflow that generates chromosome-scale assemblies, including organelles. This allows the user to rely on a single workflow to obtain the complete genomic information, including both the nuclear and organellar genomes. Colora requires PacBio HiFi and Hi-C reads as mandatory inputs, and ONT reads can be optionally integrated into the process. With Colora, it is possible to obtain a scaffolded primary assembly or a phased assembly with separate haplotypes. Colora was primarily developed for plant genome assembly, as it includes plant-specific tools such as Oatk (Zhou *et al.* 2024), but it can also be used for other organisms. Colora is the first automated workflow for *de novo* genome assembly implemented in Snakemake that integrates PacBio Hifi, ONT, and Hi-C reads to produce complete chromosome-scale assemblies.

2 Materials and methods

Colora was constructed using the Snakemake workflow manager (version $\geq$8.0.0). The whole workflow is installed and runs using virtual environments created by Conda (Grüning *et al.* 2018, Anaconda Software Distribution 2024). We provide frozen environments containing the tools with the version already tested to avoid breakage due to dependency updates.

Figure 1 shows a schematic representation of Colora.

The choice of tools integrated into the workflow is based on those chosen by world-leading biodiversity projects, such as the Earth BioGenome Project (EBP) (https://www.earthbiogenome.org/), the Darwin Tree of Life (DToL) (https://www.

Received: 10 January 2025; **Revised:** 18 March 2025; **Editorial Decision:** 10 April 2025; **Accepted:** 14 April 2025

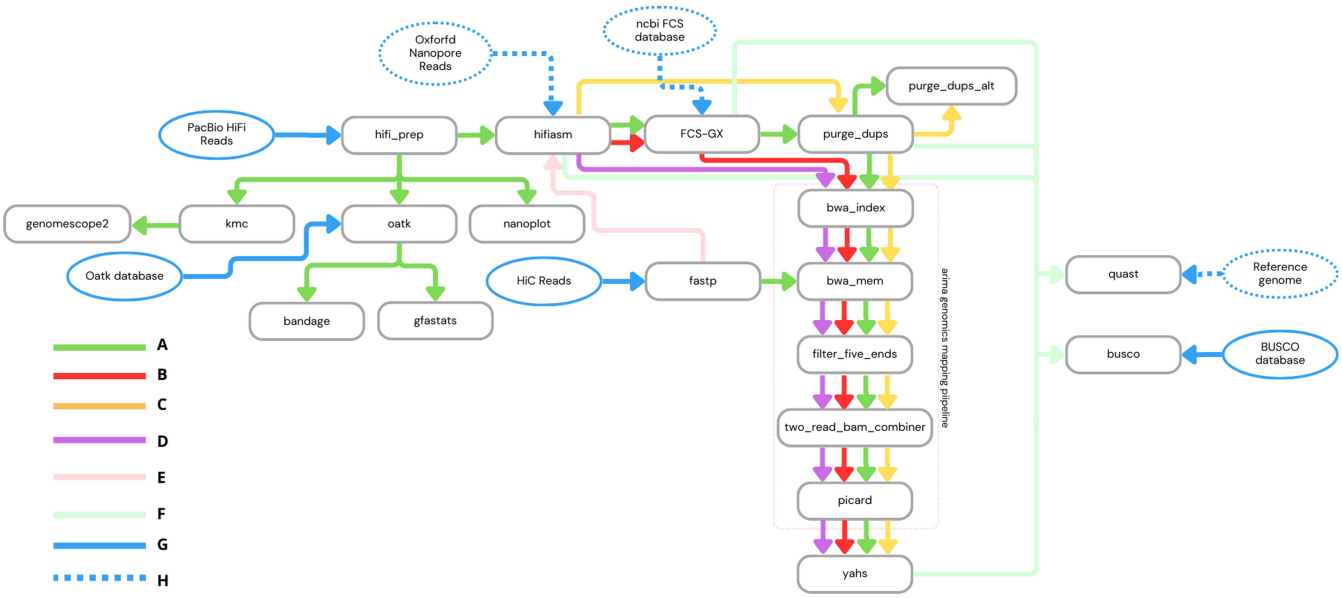

Figure 1. Flowchart illustrating the Colora Snakemake workflow. (A) Full workflow in case of primary assembly; (B) skip Purge_dups step, full workflow in case of phased assembly; (C) skip FCS-GX step; (D) skip Purge_dups and FCS-GX steps; (E) executed only in case of phased assembly; (F) assembly quality inspection (checkpoints are dependent on the workflow configuration); (G) mandatory input; (H) optional input.

darwintreeoflife.org/), and the Vertebrate Genomes Project (VGP) (https://vertebrategenomesproject.org/), and on our previous benchmarking (Obinu *et al.* 2024).

2.1 Reads pre-processing

The workflow executes the quality assessment of the reads used to perform the assembly process. After being joined in a single file with the rule `hifi_prep` if needed, PacBio HiFi reads are checked for quality with NanoPlot (De Coster *et al.* 2018). As this kind of read typically does not need to be filtered before the assembly process, we did not implement an automatic filtering step. Hi-C reads are automatically checked and filtered for quality, and adapters are removed with Fastp (Chen *et al.* 2018). ONT reads (if available) must be previously quality inspected and filtered if necessary.

2.2 K-mers analysis, organelles, and genome assembly

The k-mers are counted with KMC (Deorowicz *et al.* 2013, 2015, Kokot *et al.* 2017) from the PacBio HiFi reads, and the k-mer spectrum is plotted with GenomeScope2 (Rhyker Ranallo-Benavidez *et al.* 2020).

The organelles are assembled from the PacBio HiFi reads using Oatk (Zhou *et al.* 2024).

The genomic assembly process starts with Hifiasm. Hifiasm can be run with PacBio HiFi reads only, in which case the result will be a primary assembly, or in "Hi-C mode," i.e. integrating Hi-C reads to separate the haplotypes in heterozygous species, resulting in two separate assemblies representing the two haplotypes (hap1 and hap2). In both cases, ONT reads can be optionally integrated if available.

Table 1 outlines the differences between workflow paths that can be chosen by the user after the contig-level assembly with Hifiasm. It shows the possibility of performing the steps of contaminants removal with FCS-GX (Astashyn *et al.* 2024), removal of duplications and overlaps with Purge_dups (Guan *et al.* 2020), Hi-C reads mapping to primary contigs with the Arima Genomics pipeline (Arima Genomics 2017, https://github.com/ArimaGenomics/mapping_pipeline), and assembly scaffolding with YaHS (Zhou *et al.* 2023) depending on the chosen configuration.

The Arima Genomics Mapping Pipeline (Arima Genomics 2017) has been adapted to Snakemake from its original bash form. In addition, the flag `-M` has been added to the `bwa mem` commands, which marks shorter split hits as secondary (Li and Durbin 2009) and is recommended for compatibility with Picard (Broad Institute 2019, https://broadinstitute.github.io/picard/), which is used in later scripts of the same pipeline.

2.3 Results quality assessment

Assemblies produced at different steps along the workflow, depending on the configuration, are submitted to quality assessment. The genome assembly quality assessment is performed using BUSCO (Simão *et al.* 2015, Manni *et al.* 2021) and QUAST (Mikheenko *et al.* 2018).

The quality assessment of organelle assemblies is performed with Gfastats (Formenti *et al.* 2022) and Bandage (Wick *et al.* 2015).

Hi-C contact maps of the scaffolded assemblies were produced after the completion of the workflow to gain further evidence about the success of the assembly process. They were obtained using the code shown in Additional File S0 and visualized with PretextView (Harry and Guan 2024, https://github.com/sanger-tol/PretextView).

2.4 User specifications

The user can specify the desired parameters to use along the workflow and the route to follow by editing the `config.yaml` file. This is essential for a successful run of Colora.

Before starting the workflow, the user must be sure to have all the mandatory inputs, which are PacBio Hifi and Hi-C reads, the Oatk database (Zhou *et al.* 2024), and the BUSCO database for the species studied (light blue continuous lines in Fig. 1). Depending on the chosen configuration, other optional inputs that must be available before starting are ONT

Table 1. Comparison of possible workflow paths following the primary contigs assembly with Hifiasm.[a]

Workflow path	Contaminants removal from primary contigs	Overlap and duplications removal	Hi-C reads mapping to primary contigs	Assembly scaffolding	Applicable for primary assembly	Applicable for phased assembly	Arrow colour in Fig. 1
Skip contaminants, overlaps, and duplications removal	×	×	✓	✓	✓	✓	Purple
Skip overlaps and duplications removal	✓	×	✓	✓	✓	✓	Red
Skip contaminants removal	×	✓	✓	✓	✓	×	Yellow
Full workflow	✓	✓	✓	✓	✓	×	Green

[a] Symbols: ✓ for "yes," × for "no."

reads, the NCBI FCS-GX database, and the reference genome, optionally with the related annotation GFF file, for the species studied (light blue dotted lines in Fig. 1). These inputs can be specified through the `config.yaml` file, as explained in detail in `config/README.md` and in the examples on our GitHub repository wiki (https://github.com/LiaOb21/colora).

The `config.yaml` allows the users to change several parameters for customization of the tools according to their needs for the studied species. In the `config/README.md`, we have highlighted the key parameters that a user should check in the `config.yaml` and have provided links to the upstream documentation where users may find further details for each tool.

2.5 Testing of the workflow

To test the pipeline, we used three publicly available datasets:

1) *Rhizophagus irregularis*, strain G1, was obtained from the BioProject PRJNA922099; the dataset included PacBio HiFi and Hi-C reads. This organism is a heterokaryotic arbuscular mycorrhizal fungus, in which cells two different kinds of haploid nuclei coexist (Sperschneider *et al.* 2023). The size of the reference genome is 146.8 Mbp, and it is organized into 32 chromosomes (Manley *et al.* 2023).
2) For *Arabidopsis thaliana*, the BioProject PRJCA005809 dataset included PacBio HiFi, ONT, and Hi-C reads (Wang *et al.* 2022). This organism is the model plant for genomics, and its genome size is 135 Mbp organized in 5 chromosomes. The size of the reference genome TAIR10.1 is 119.1 Mbp (Kaul *et al.* 2000). Since this species is autogamous, the two haplotypes are identical by descendent (IBD), and its genome is therefore represented by a single haplotype.
3) *Malus domestica*, cultivar Fuji, was obtained from BioProject PRJNA814760; the dataset included PacBio HiFi and Hi-C reads (Li *et al.* 2024a). This is an agronomically important species, and the size of the reference genomes is 703 Mbp organized in 17 chromosomes. This species is allogamous and heterozygous, and its genome is therefore expected to be diploid.

These datasets were used to produce *de novo* genome assemblies in previous studies. The assemblies produced in the original studies and the reference genome for each species were quality inspected using QUAST and BUSCO to compare the results with the quality of the assemblies produced with Colora.

For the *R. irregularis* and *M. domestica* datasets, the reads were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) using SRA tools (National Center for Biotechnology Information 2024, https://github.com/ncbi/sra-tools). *Arabidopsis thaliana* data were downloaded from NGDC (https://ngdc.cncb.ac.cn/).

The codes used to test the workflow are available in the wiki on our GitHub repository.

In addition, we provide a test dataset for users to test the workflow within a limited amount of time. This dataset is a subset of reads of *Saccharomyces cerevisiae* obtained from the BioProject PRJNA1075684 for PacBio and ONT reads and from the BioProject PRJNA1013711 for Hi-C reads (Li *et al.* 2024b). This dataset is for testing purposes only.

Before running Colora using the *A. thaliana* dataset, we inspected ONT reads with NanoPlot (De Coster *et al.* 2018) to assess read length and quality distributions. To improve data quality, the reads were filtered using NanoFilt (De Coster *et al.* 2018) with the parameter `-l 500` to exclude reads shorter than 500 bp. Given the limited mean read length of the dataset, we applied a shallow filter to avoid over-exclusion of data.

All the other parameters used to produce the assemblies are shown in the relative `config.yaml` files in Additional Files (S1–S3).

Colora was tested on two different HPC systems: (i) a cluster with up to 576 cores and 18 TB of shared main memory, which was used to obtain the results presented in this work, and (ii) a separate HPC system from which a single node was used, equipped with 36 cores (2 × Intel Xeon E5-2695, Broadwell @ 2.1 GHz) and 256 GB of memory. On HPC systems, pipeline jobs were executed via SLURM. Additionally, Colora was also tested on smaller datasets using personal computers with CPUs Intel-i7, 16 GB of RAM, and Ubuntu-based OS.

While testing, we monitored the memory consumption along the workflow to assess performance and provide users with suggestions for memory requirements. These data are reported in the `config/README.md` file and in the `config.yaml` files in Additional Files (S1–S3).

After the successful workflow, we produced a Snakemake report using the command `snakemake–report` to monitor the runtime of each rule.

3 Results

3.1 Workflow performance

Figure 2 shows the runtime graphs obtained from the Snakemake report for the three species. The Hifiasm rule was the most demanding rule in terms of time for the three datasets. Hifiasm required the highest amount of time for the *M. domestica* dataset.

Figure 3 shows the maximum resident size (kbytes) for each rule measured for the three datasets, i.e. the highest amount of RAM each rule required while it was running.

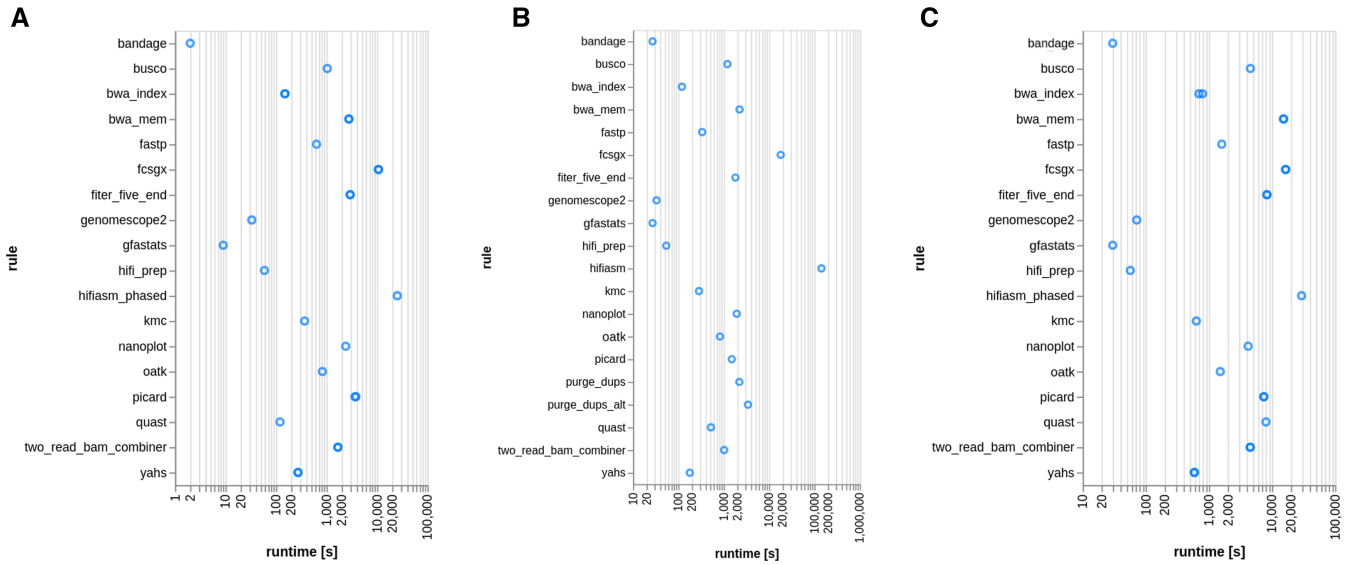

Figure 2. Runtime (seconds) required by each rule for *R. irregularis* (A), *A. thaliana* (B), and *M. domestica* (C).

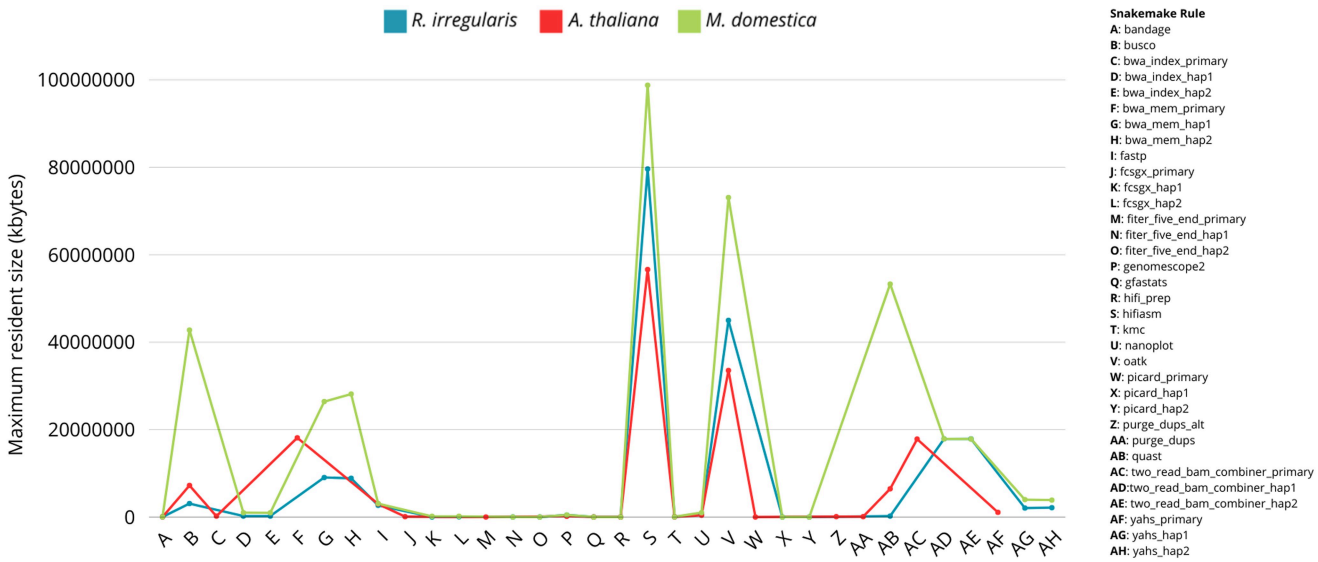

Figure 3. Maximum resident size (kbytes) required by each rule for the three datasets.

Hifiasm was the most demanding rule in terms of memory for the three datasets, and it required the highest amount of RAM for the *M. domestica* dataset.

This is in line with the expectations, as the *M. domestica* genome was the largest among those analyzed. Overall, the runtime and RAM requirements for each rule are dependent on the size of the dataset. Therefore, the workflow performance is strictly dependent on the dataset analyzed.

3.2 Reads quality assessment

The complete NanoPlot reports for PacBio HiFi reads (S4–S6), Fastp reports for Hi-C reads (S7–S9) for the three datasets, and the NanoPlot report for ONT *A. thaliana* dataset (S10) are shown in Additional Files.

Table 2 summarizes the read characteristics of the *R. irregularis*, *A. thaliana*, and *M. domestica* datasets.

3.3 GenomeScope2 profiles

For *R. irregularis* (Fig. 4A), GenomeScope2 calculated an estimated genome size of 142 701 404 bp, which is comparable to the size of the reference genome ASM2621079v1, and the genome size obtained in the original paper (Additional File S11). It was possible to obtain a similar estimate from the k-mer analysis only using the parameter `-p 1` for GenomeScope2, which is used for homozygous genomes. This species is known to be heterokaryotic, i.e. it carries two different types of haploid nuclei in its cells (Sperschneider *et al.* 2023). It is likely that the small peak observed at coverage ∼50, erroneously placed by the model among the errors, represents the second haploid nucleus.

GenomeScope2 estimated a genome size of 120 320 924 bp for *A. thaliana* (Fig. 4B), which is comparable to the reference genome of the species TAIR10.1, while the genome size of the assembly obtained in the original paper was 133.7 Mbp.

Table 2. Overview of the read characteristics of the *R. irregularis* dataset (BioProject PRJNA922099), *A. thaliana* (BioProject PRJCA005809), and *M. domestica* (BioProject PRJNA814760).

Species	Dataset	Mean read length	Mean read quality	Number of reads	Total bases	Coverage
Rhizophagus irregularis	PacBio HiFi	12 812.1	21.0	2 341 873	30 004 306 961	204.11
	Hi-C	150	>30	236 467 312	35 470 097 000	241.29
Arabidopsis thaliana	PacBio HiFi	15 094.4	27.1	1 517 433	22 904 700 074	169.66
	Hi-C	150	>30	140 957 500	21 143 625 000	156.62
	ONT	**18 541.3**	**11.1**	**3 064 191**	**56 814 196 989**	**420.84**
Malus domestica	PacBio HiFi	11 749.8	27.2	4 523 532	53 150 548 964	75.9
	Hi-C	148	>30	676 416 866	100 522 162 000	143.60

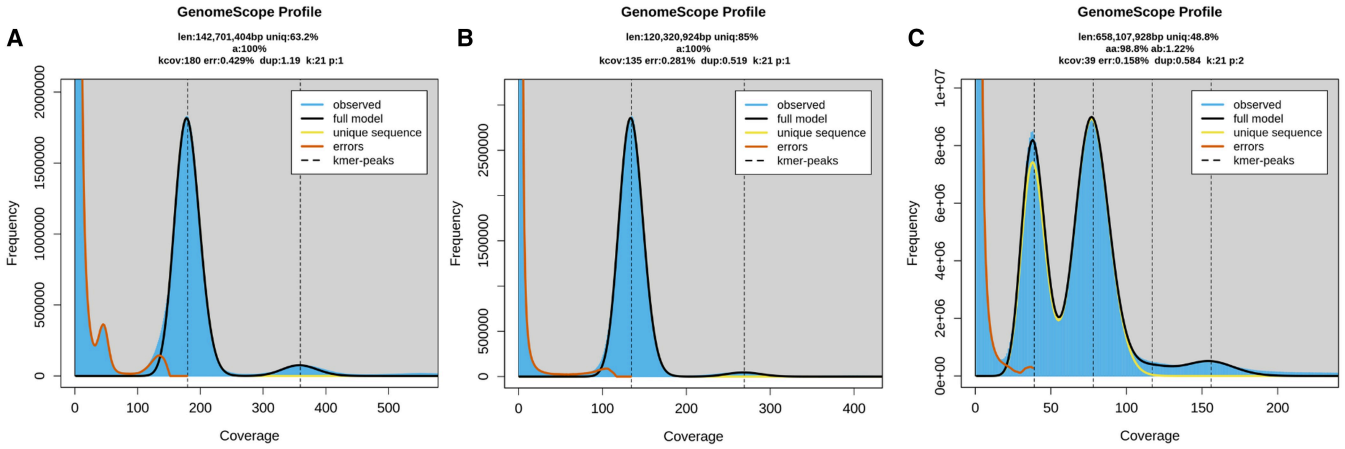

Figure 4. GenomeScope2 k-mer profiles for *R. irregularis* (A), *A. thaliana* (B), and *M. domestica* (C).

As expected, the k-mer spectrum fits the haploid genome model, showing only one peak.

For *M. domestica*, GenomeScope2 estimated genome size of 658 107 928 bp (Fig. 4C), close in size to the haploid assemblies from the original paper but lower compared to the reference genome ASM211411v1 and to the consensus haploid assembly from the original paper (Additional File S11). The k-mer spectrum was consistent with what is expected for a heterozygous genome, showing two peaks.

3.4 Organelle assembly quality assessment

Figure 5 shows the statistics obtained with Gfastats (Formenti *et al.* 2022) and the plot obtained with Bandage (Wick *et al.* 2015) for the mitochondrial assembly of the three species. For all the datasets, the mitochondrial assemblies were composed of a single, gapless contig and segment. Compared to the reference mitochondrion, the assemblies displayed close concordance in length.

Figure 6 displays the statistics generated using Gfastats and the visualization created with Bandage for the chloroplast assemblies of *A. thaliana* and *M. domestica*. In both datasets, the chloroplast assemblies consisted of a single, gapless contig and segment. The assembly lengths were compared to the reference chloroplast length, revealing comparable results.

Complete Gfastats results for mitochondria (S12–S14) and chloroplast (S15 and S16) are shown in Additional Files.

3.5 Genome assembly quality assessment

3.5.1 Rhizophagus irregularis

The number of contigs for the two haplotypes decreased along the workflow from 183 to 152 for hap1 and from 109 to 97 for hap2 due to the process of decontamination and scaffolding (Fig. 7).

The number of contigs in the reference genome and the assemblies obtained in the original paper is lower compared to our assemblies (Additional File S11). However, these assemblies were manually curated (Manley *et al.* 2023, Sperschneider *et al.* 2023). For our assemblies, further improvements may be achieved with manual curation, as shown by the Hi-C contact maps (Additional File S17A and B), which displayed the presence of small contigs not included in the chromosome-length scaffolds for both assemblies. The assemblies could be improved further by removing contaminations from mitochondrial sequences, which are not automatically removed by FCS-GX.

The total length of the final assembly was 149.3 Mbp for hap1 and 146.5 Mbp for hap2, which is close to the genome size predicted by GenomeScope2 and comparable to the total length obtained for the assemblies of the original paper and the reference genome (Additional File S11).

The N50 of the final assembly was 4.9 Mbp for both haplotypes, which is comparable to the results obtained for the assemblies of the original paper and the reference genome. The L50 was 14 for both haplotypes in the final scaffolded assembly.

The complete QUAST report for *R. irregularis* is shown in Additional File (S11).

Rhizophagus irregularis assemblies were compared with the BUSCO lineage mucoromycota_odb10 (Fig. 8A). The BUSCO scores for the final hap1 assembly were 98.2% complete BUSCOs (96.3% single copy, 1.9% duplicated), 0.3% fragmented BUSCOs, and 1.5% missing BUSCOs. For hap2,

Species	*R. irregularris (A)*	*A. thaliana (B)*	*M. domestica (C)*
Contigs	1	1	1
Segments	1	1	1
Gaps	0	0	0
Total length (bp)	70,272	367,810	396,948
Reference length (bp)	70,800	367,808	396,947

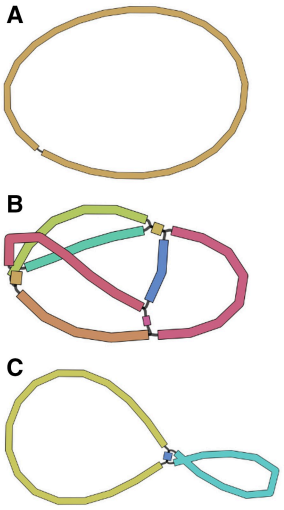

Figure 5. Gfastats statistics and Bandage visualization for *R. irregularis* (A), *A. thaliana* (B), and *M. domestica* (C) mitochondria. NCBI reference sequences (RefSeq) are NC_014489.1, NC_037304.1, and NC_018554.1 respectively.

Species	*A. thaliana (A)*	*M. domestica (B)*
Contigs	1	1
Segments	1	1
Gaps	0	0
Total length (bp)	154,479	160,072
Reference length (bp)	154,478	160,288

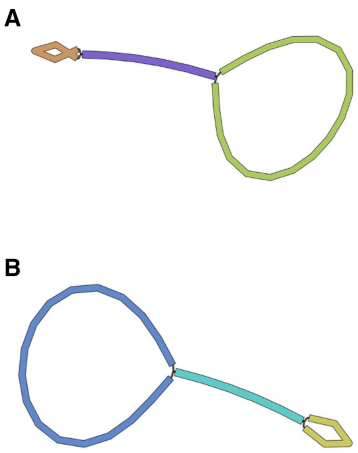

Figure 6. Chloroplast gfastats statistics and Bandage visualization for *A. thaliana* (A), and *M. domestica* (B). NCBI reference sequences (RefSeq) are NC_000932.1 and NC_061549.1 respectively.

the results showed 97.8% complete BUSCOs, (97.0% single copy and 0.8% duplicated), 0.4% fragmented BUSCOs, and 1.8% missing BUSCOs. Overall, the results show that BUSCO scores remained unaltered throughout the workflow. The BUSCO metrics for the reference genome and the assemblies published in the original paper were comparable to those obtained for the assemblies generated with Colora (Additional File S18).

3.5.2 Arabidopsis thaliana

The number of contigs was drastically reduced during the workflow, from the initial 507 to the final 18, especially due to the removal of haplotigs and overlaps by Purge_dups (Fig. 9).

The total length of the final assembly was 134.38 Mbp. This result is distant from the estimated genome size obtained with GenomeScope2, but it is in line with the results obtained in the original paper, where the authors obtained a genome assembly of 133.7 Mbp. The reference genome TAIR10.1 has a genome size of 119.1 Mbp.

The L50 was equal to 2, indicating the possible introduction of a false join during the scaffolding process since we would expect this value to be equal to 3. The false join between the first two chromosomes was clearly shown by the Hi-C contact map (Additional File S17C).

The N50 for the final assembly was 32.6 Mbp, which is higher compared to the results obtained for the reference genome and the original paper due to the presence of the false join between two scaffolds.

For this species, we expect 5 chromosomes, and this result can be achieved with manual curation of the assembly and by removing contaminations from mitochondrial and chloroplast sequences.

The complete QUAST report for *A. thaliana* is shown in Additional Files (S11).

Arabidopsis thaliana assemblies were compared with the BUSCO lineage brassicales_odb10 (Fig. 8B). The BUSCO scores for the final assembly are 100.0% of complete BUSCOs (99.1% single copy and 0.9% duplicated), 0.1%

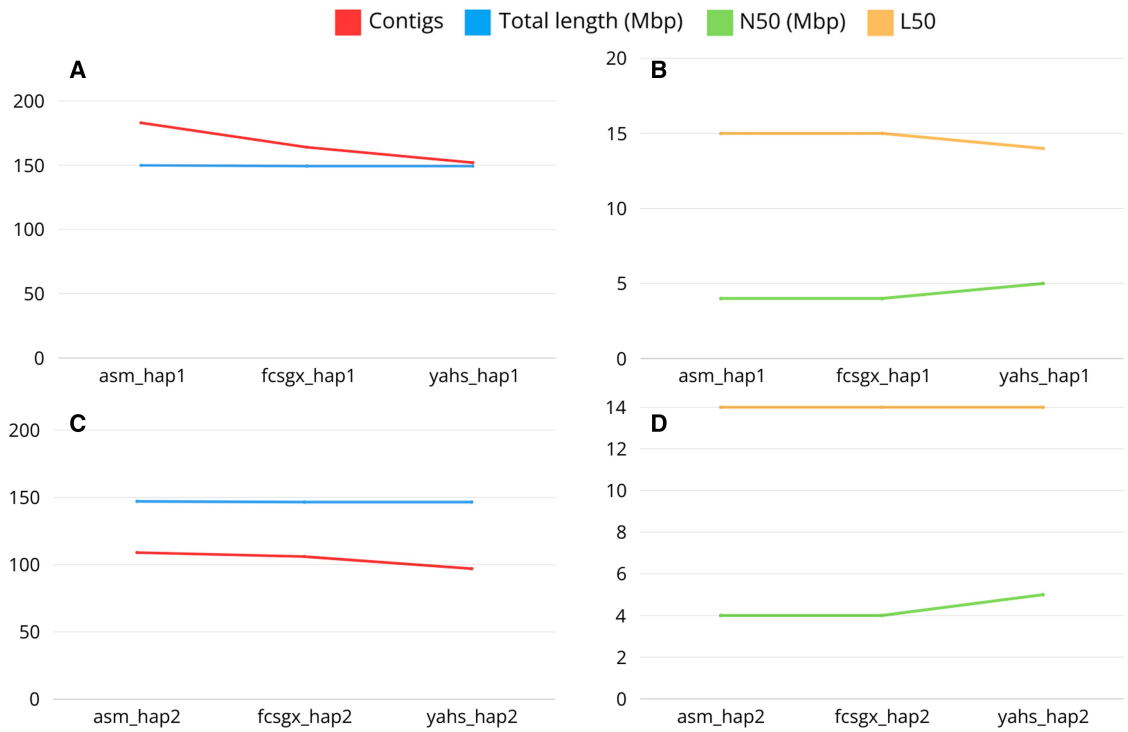

Figure 7. Quast metrics for *R. irregularis* for haplotype 1 (A, B) and haplotype 2 (C, D).

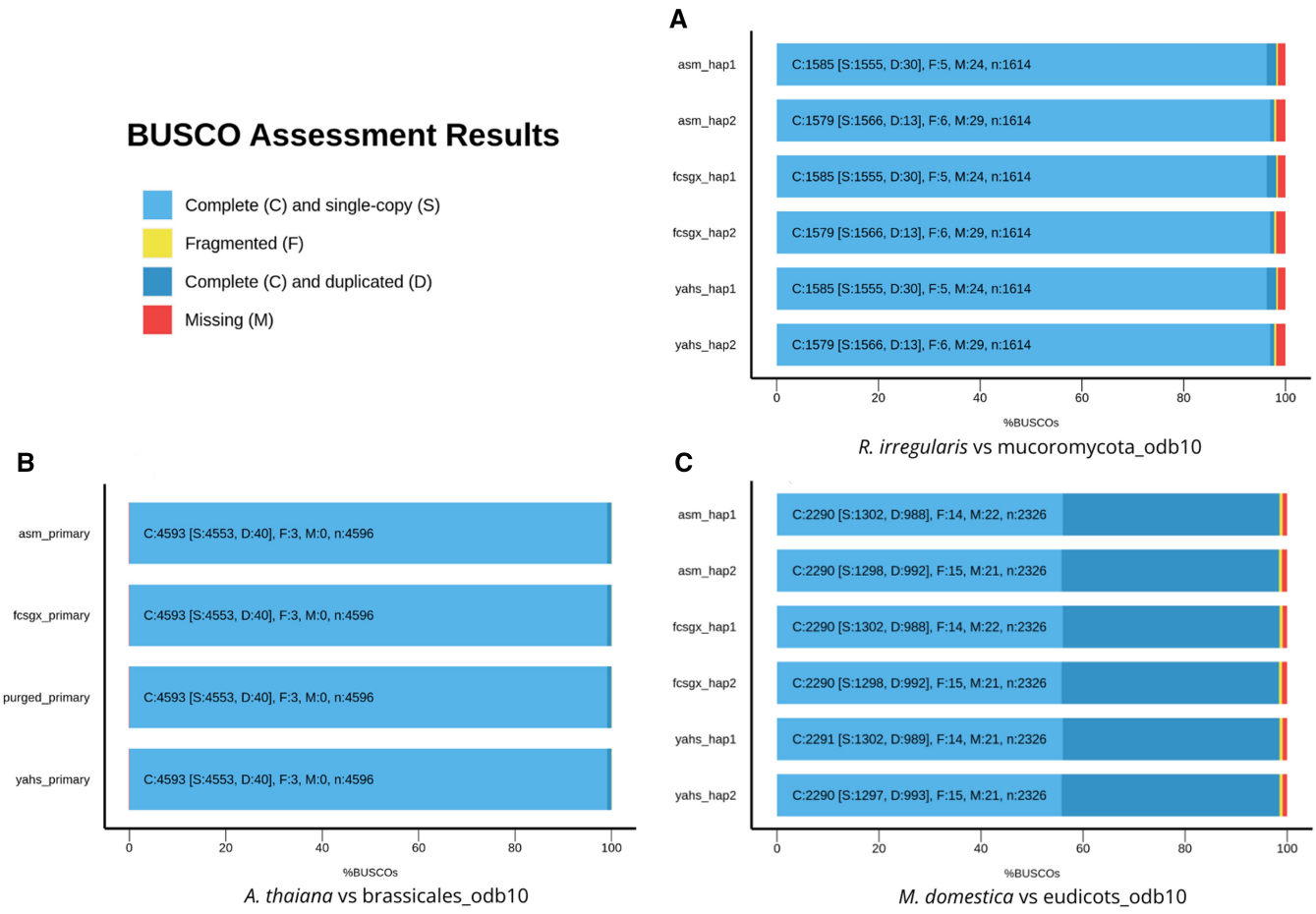

Figure 8. BUSCO Assessment results for *R. irregularis* (A), *A. thaliana* (B), and *M. domestica* (C).

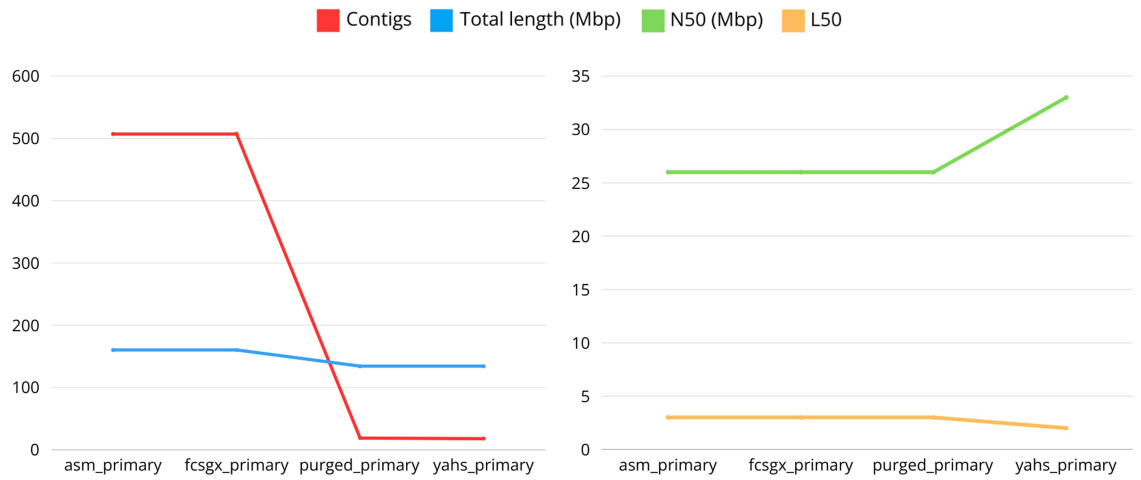

Figure 9. Quast metrics for *A. thaliana*.

fragmented BUSCOs, and no missing BUSCOs. Overall, the results show that BUSCO scores remained unaltered along the workflow. The BUSCO metrics for the assemblies obtained with Colora were similar to those observed for the reference genomes and the assembly published in the original paper (Additional File S18).

3.5.3 Malus domestica

The number of contigs decreased along the workflow for both haplotypes. However, they did not reach a number of contigs matching the expected number of chromosomes for this species, which is 17 (Fig. 10). The same is true for the number of contigs observed in the reference genome and the assemblies produced in the original paper (Additional File S11). However, the assemblies can be improved with further manual curation and by removing contaminations from mitochondrial sequences.

The total length of the final assembly was 686.8 Mbp and 662.6 Mbp for hap1 and hap2, respectively. These results are comparable with the expected genome size shown by Genomescope2 and with the size obtained in the original paper for the two haplotypes. The reference genome ASM211411v1 for this species is 703 Mbp, while the consensus assembly obtained in the original paper showed a length equal to 736.9 Mbp (Additional Files S11).

The N50 was 37.2 Mbp and 36.9 Mbp for hap1 and hap2, respectively, showing values comparable to the assemblies from the original paper and to the reference genome (Additional Files S11).

The L50 was equal to 7 for both the haplotypes. For this species, we would expect an L50 = 9; therefore, this is an indication of the presence of false joins introduced during the process of scaffolding, which was confirmed by the Hi-C contact maps (Additional File S17D and E).

Malus domestica assemblies were compared with the BUSCO lineage eudicots_odb10 (Fig. 8C). The final BUSCO scores for hap1 were 98.5% complete BUSCOs (56.0% single copy and 42.5% duplicated), 0.6% fragmented BUSCOs, and 0.9% missing BUSCOs. The final BUSCO scores for hap2 were 98.5% complete BUSCOs (55.8% single copy and 42.7% duplicated), 0.6% fragmented BUSCOs, and 0.9% missing BUSCOs. For hap1, complete and duplicated BUSCOs slightly increased along the workflow, and missing BUSCOs slightly decreased. For hap2, complete and single copy BUSCOs slightly decreased along the workflow, while complete and duplicated BUSCOs slightly increased. The high degree of duplication in *M. domestica* genome is known and is the result of small-scale and large-scale whole-genome duplication events during its evolutionary history (Sanzol 2010, Li *et al.* 2019). The BUSCO metrics for the assemblies obtained with Colora were comparable to those observed for the reference genomes and the assemblies published in the original paper (Additional File S18).

4 Discussion

Colora is the first Snakemake workflow for *de novo* genome assembly that combines PacBio Hifi, ONT, and Hi-C reads to obtain complete chromosome-scale assemblies. A previously published Snakemake workflow for *de novo* genome assembly, SnakeCube, utilizes MinION ONT and Illumina reads as input (Angelova *et al.* 2022). However, there are Nextflow and Galaxy pipelines that share part of the assembly strategy and tools with Colora, and which inspired Colora. These pipelines are the Earth Biogenome assembly pipeline (National Bioinformatics Infrastructure Sweden (NBIS) 2024, https://github.com/NBISweden/Earth-Biogenome-Project-pilot) and the DToL assembly pipeline (Krasheninnikova *et al.* 2024, https://zenodo.org/records/10990898), which are implemented in Nextflow, and the VGP pipeline, which is implemented in Galaxy (Larivière *et al.* 2024). Colora has many aspects in common with these pipelines but also its peculiarities. For example, the decontamination process is performed using FCS-GX and applied to the contig-level assembly. FCS-GX has only recently been implemented in the Earth BioGenome assembly pipeline. In Colora, this step is optional: this allows the user to use the pipeline in small systems, as the FCS-GX database needs a substantial amount of disk space, and the process requires a large amount of RAM to run. The removal of haplotigs with Purge_dups is also optional (in the case of primary assembly), as we observed that in large and highly repetitive genomes this can lead to a substantial reduction in complete BUSCOs (unpublished work). The tool used to assemble the organelles is Oatk, which is optimized for the assembly of complex plant organelles, but it can also be used for simple organelle genomes. Colora has the peculiarity of evaluating organelle assemblies using Gfastats and Bandage.

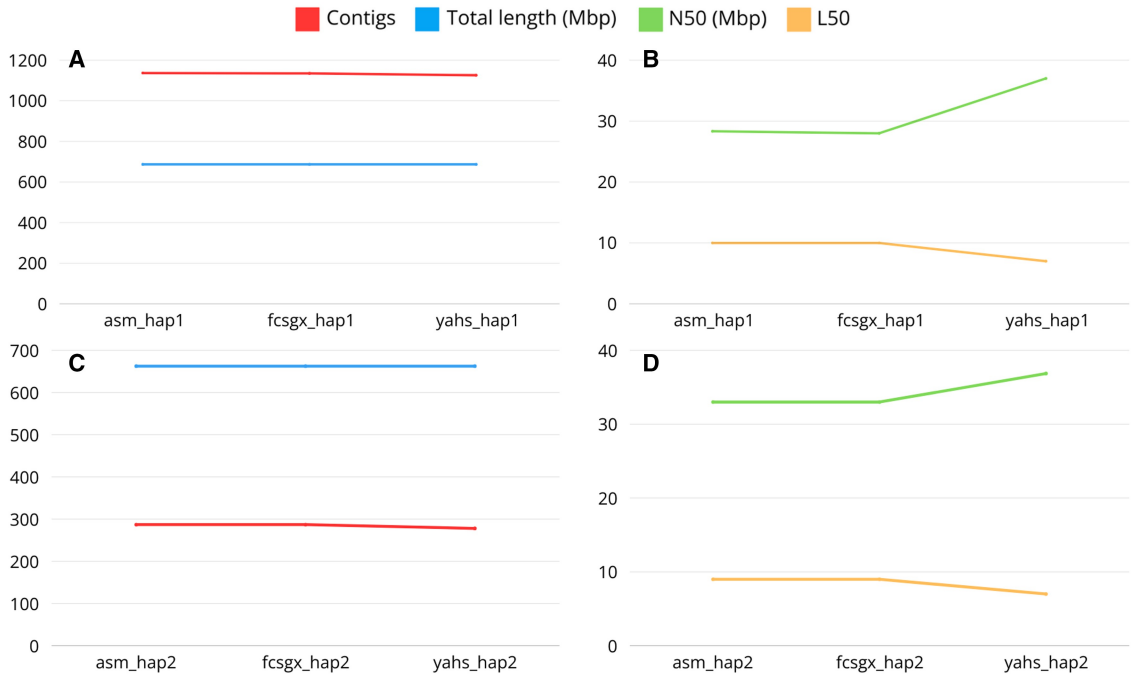

Figure 10. Quast metrics for *M. domestica* haplotype 1 (A, B) and haplotype 2 (C, D).

Overall, Colora automatically produces high-quality *de novo* assemblies, which are comparable to the assemblies obtained in the original papers with a step-by-step workflow. Subsequent manual curation, as suggested by Howe *et al.* (2021), is necessary to further improve the quality of the assemblies.

5 Limitations and future work

Colora is set up to run on Linux systems. The version described in this paper is v_1.0.0. At the moment, Colora does not handle genomes with ploidy levels greater than two. In the future, we aim to incorporate more tools for genome quality assessment, such as Merqury for genome quality assessment based on k-mer (Rhie *et al.* 2020), and Inspector for structural correctness (Chen *et al.* 2021). We plan to add a method that automatically removes contigs identified as organelles from the nuclear genome. We aim to keep all the tools included in the pipeline up to date to their latest versions, and to implement the containerization with Docker and Singularity.

6 Conclusions

Colora is a reliable, state-of-the-art, and portable pipeline for automated *de novo* genome assembly, specifically designed with plant genomes in mind. The primary goal of the workflow is to be user-friendly and adaptable to various systems and datasets. With detailed documentation to support users, we believe it will be valuable for researchers worldwide seeking a customizable, automated, and easy-to-implement solution for obtaining high-quality genome assemblies.

Acknowledgements

The authors would like to thank the organizers of Biodiversity Genomics Academy 2023 (https://bga23.org/), which was essential for acquiring the knowledge that determined the success of this project.

Author contributions

Lia Obinu (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [lead], Software [lead], Validation [lead], Writing—original draft [lead]), Timothy G. Booth (Conceptualization [supporting], Methodology [equal], Software [equal], Supervision [equal], Writing—review & editing [equal]), Heleen De Weerd (Conceptualization [supporting], Methodology [supporting], Software [supporting], Validation [equal], Writing—review & editing [equal]), Urmi Trivedi (Conceptualization [supporting], Methodology [supporting], Supervision [lead], Writing—review & editing [equal]), and Andrea Porceddu (Funding acquisition [lead], Supervision [supporting], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This study includes part of a PhD project carried out by L.O. at the PhD School of Agricultural Sciences of the University of Sassari, funded by the University of Sassari. L.O. was also supported by the ERASMUS for Traineeship program awarded to the University of Sassari. This study was carried out within the Agritech National Research Centre and received funding from the European Union Next-Generation EU [PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D. D. 1032 17/06/2022, CN00000022]. This manuscript reflects only the authors' views and opinions and neither the European

Union nor the European Commission can be considered responsible for them.

Data availability

Project name: Colora
Project home page: https://github.com/LiaOb21/colora
Operating system(s): Linux
Programming language: Python, Bash, Perl, Awk
Other requirements: Miniconda, Snakemake
License: MIT

Colora is also available at the Snakemake Workflow Catalog (https://snakemake.github.io/snakemake-workflow-catalog/? usage=LiaOb21%2Fcolora) and the code has been deposited in Zenodo under DOI https://doi.org/10.5281/zenodo.13321576.

The public raw data used to test the workflow are available from the NCBI SRA database for *R. irregularis* (https://www.ncbi.nlm.nih.gov/sra? LinkName=biosample_sra&from_uid=32643414) and *M. domestica* (https://www.ncbi.nlm.nih.gov/sra? LinkName=biosample_sra&from_uid=26566345) and from the NGDC for *A. thaliana* (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA005809).

References

Anaconda Software Distribution. Conda. 2024.

Angelova N, Danis T, Lagnel J *et al.* Snakecube: containerized and automated pipeline for de novo genome assembly in HPC environments. *BMC Res Notes* 2022;**15**:98. https://doi.org/10.1186/S13104-022-05978-5/FIGURES/2

Arima Genomics. Mapping pipeline for data generated using Arima-HiC. 2017.

Astashyn A, Tvedte ES, Sweeney D *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* 2024;**25**:60. https://doi.org/10.1186/S13059-024-03198-7

Broad Institute. Picard toolkit. 2019.

Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/BIOINFORMATICS/BTY560

Chen Y, Zhang Y, Wang AY *et al.* Accurate long-read de novo assembly evaluation with inspector. *Genome Biol* 2021;**22**:312, 12. https://doi.org/10.1186/s13059-021-02527-4

De Coster W, D'Hert S, Schultz DT *et al.* Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;**34**:2666–9. https://doi.org/10.1093/bioinformatics/bty149

Deorowicz S, Debudaj-Grabysz A, Grabowski S. Disk-based k-mer counting on a pc. *BMC Bioinformatics* 2013;**14**:160–12. https://doi.org/10.1186/1471-2105-14-160/FIGURES/8

Deorowicz S, Kokot M, Grabowski S *et al.* Kmc 2: fast and resource-frugal k-mer counting. *Bioinformatics* 2015;**31**:1569–76. https://doi.org/10.1093/BIOINFORMATICS/BTV022

Formenti G, Abueg L, Brajuka A *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* 2022;**38**:4214–6. https://doi.org/10.1093/BIOINFORMATICS/BTAC460

Grüning B, Dale R, Sjödin A *et al.* The Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–6. https://doi.org/10.1038/s41592-018-0046-7

Guan D, Guan D, McCarthy SA *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics (Oxford, England)* 2020;**36**:2896–8. https://doi.org/10.1093/BIOINFORMATICS/BTAA025

Guan D, McCarthy SA, Ning Z *et al.* Efficient iterative hi-c scaffolder based on n-best neighbors. *BMC Bioinformatics* 2021;**22**:612–6. https://doi.org/10.1186/s12859-021-04453-5

Harry E, Guan S. *PretextView: Paired REad TEXTure Viewer*. 2024.

Howe K, Chow W, Collins J *et al.* Significantly improving the quality of genome assemblies through curation. *Gigascience* 2021;**10**:1–9. https://doi.org/10.1093/GIGASCIENCE/GIAA153

Jayakumar V, Sakakibara Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform* 2019;**20**:866–76. https://doi.org/10.1093/bib/bbx147

Jung H, Winefield C, Bombarely A *et al.* Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends Plant Sci* 2019;**24**:700–24. https://doi.org/10.1016/J.TPLANTS.2019.05.003

Kaul S, Koo HL, Jenkins J *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;**408**:796–815. https://doi.org/10.1038/35048692

Kokot M, Dlugosz M, Deorowicz S. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics* 2017;**33**:2759–61. https://doi.org/10.1093/BIOINFORMATICS/BTX304

Krasheninnikova K, Qi G, Muffato M *et al. sanger-tol/genomeassembly: v0.10.0 – hideous zippleback*, 2024.

Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLoS One* 2017;**12**:e0177459. https://doi.org/10.1371/JOURNAL.PONE.0177459

Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2. https://doi.org/10.1093/BIOINFORMATICS/BTS480

Larivière D, Abueg L, Brajuka N *et al.* Scalable, accessible and reproducible reference genome assembly and evaluation in galaxy. *Nat Biotechnol* 2024;**42**:367–70. https://doi.org/10.1038/s41587-023-02100-3

Li H, Huang C-H, Ma H. Whole-Genome Duplications in Pear and Apple. In: Korban S (ed.), *The Pear Genome. Compendium of Plant Genomes*. Cham: Springer, 2019, 279–99. https://doi.org/10.1007/978-3-030-11048-2_15

Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324

Li W, Chu C, Li H *et al.* Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nat Genet* 2024a;**56**:505–16. https://doi.org/10.1038/s41588-024-01657-2

Li Y, Lee J, Bai L. DNA methylation-based high-resolution mapping of long-distance chromosomal interactions in nucleosome-depleted regions. *Nat Commun* 2024b;**15**:4358. https://doi.org/10.1038/S41467-024-48718-Y

Manley BF, Lotharukpong JS, Barrera-Redondo J *et al.* A highly contiguous genome assembly reveals sources of genomic novelty in the symbiotic fungus *Rhizophagus irregularis*. *G3 (Bethesda)* 2023;**13**:jkad077. https://doi.org/10.1093/g3journal/jkad077

Manni M, Berkeley MR, Seppey M *et al.* Busco update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**:4647–54. https://doi.org/10.1093/MOLBEV/MSAB199

Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux J* 2014;**239**:2.

Mikheenko A, Prjibelski A, Saveliev V *et al.* Versatile genome assembly evaluation with quast-lg. *Bioinformatics* 2018;**34**:i142–50. https://doi.org/10.1093/bioinformatics/bty266

National Bioinformatics Infrastructure Sweden (NBIS). *Earth Biogenome Project – Pilot Workflow*. 2024.

National Center for Biotechnology Information. *The SRA Toolkit*. 2024.

Obinu L, Trivedi U, Porceddu A. Benchmarking of Hi-C tools for scaffolding plant genomes obtained from pacbio hifi and ont reads. *Front Bioinform* 2024;**4**:1462923. https://doi.org/10.3389/fbinf.2024.1462923

Rhyker Ranallo-Benavidez T, Jaron KS, Schatz MC. Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes.

Nat Commun 2020;**11**:1432–10. https://doi.org/10.1038/s41467-020-14998-3

Rhie A, Walenz BP, Koren S *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 2020;**21**:245–27. https://doi.org/10.1186/S13059-020-02134-9/FIGURES/6

Sanzol J. Dating and functional characterization of duplicated genes in the apple (*Malus domestica* borkh.) by analyzing EST data. *BMC Plant Biol* 2010;**10**:87. https://doi.org/10.1186/1471-2229-10-87

Simão FA, Waterhouse RM, Ioannidis P *et al.* Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2. https://doi.org/10.1093/bioinformatics/btv351

Sperschneider J, Yildirir G, Rizzi YS *et al.* Arbuscular mycorrhizal fungi heterokaryons have two nuclear populations with distinct roles in host–plant interactions. *Nat Microbiol* 2023;**8**:2142–53. https://doi.org/10.1038/s41564-023-01495-8

Wang B, Yang X, Jia Y *et al.* High-quality *Arabidopsis thaliana* genome assembly with nanopore and HIFI long reads. *Genomics Proteomics Bioinf* 2022;**20**:4–13. https://doi.org/10.1016/J.GPB.2021.08.003

Wick RR, Schultz MB, Zobel J *et al.* Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;**31**:3350–2. https://doi.org/10.1093/BIOINFORMATICS/BTV383

Zhou C, McCarthy SA, Durbin R. Yahs: yet another hi-c scaffolding tool. *Bioinformatics* 2023;**39**:btac808. https://doi.org/10.1093/bioinformatics/btac808

Zhou C, Brown M, Blaxter M *et al.*; The Darwin Tree of Life Project Consortium. Oatk: a de novo assembly tool for complex plant organelle genomes. bioRxiv, https://doi.org/10.1101/2024.10.23.619857, 2024, preprint: not peer reviewed.

Bioinformatics, 2025, 41, 1–11
https://doi.org/10.1093/bioinformatics/btaf175
Original Paper