



Journal homepage: <https://trpubonline.org/index.php/Biosynth>

## CASE STUDY

# Bioinformatic Pipeline Development for Long-Read Sequencing Assembly of Plant Genomes

Zsolia Novak<sup>1\*</sup>, Joao Mendes<sup>2</sup> and Katrin Peeters<sup>3</sup>

<sup>1</sup>Department of Environmental and Agricultural Sciences, Nordfjord University Center for Marine Biosciences, Norway

<sup>2</sup>Department of Microbiology and Immunology, Alpine Institute of Biomedical Sciences, Austria

<sup>3</sup>Department of Microbiology and Immunology, Nordfjord University Center for Marine Biosciences, Norway

\*Corresponding author: Zsolia Novak, Department of Environmental and Agricultural Sciences, Nordfjord University Center for Marine Biosciences, Norway; Email: [z.novak@nucmb.no](mailto:z.novak@nucmb.no)

## ABSTRACT

Reference-quality genome assembly of non-model crop and wild-relative plant species remains disproportionately difficult relative to well-resourced model organisms, owing to a combination of large genome size, high repeat content, and residual heterozygosity in outcrossing or wild-collected accessions, all of which complicate assembly from long-read sequencing data even as the underlying sequencing technology has matured. We present GRASSEMBLE, a modular bioinformatic pipeline integrating PacBio HiFi read correction, Hi-C scaffolding, automated haplotype purging and a multi-tool consensus polishing stage, and evaluate it through a case study assembly of the grass pea (*Lathyrus sativus*), a drought- and flood-tolerant legume of growing breeding interest that has lacked a chromosome-scale reference genome. Applied to 42x HiFi coverage and 38x Hi-C coverage generated from a single accession, GRASSEMBLE produced a 6.61 Gb assembly with contig N50 of 18.4 Mb, scaffolded into seven pseudomolecules corresponding to the seven grass pea chromosomes and capturing 96.8% of the estimated genome size. BUSCO completeness against the fabales\_odb10 lineage dataset reached 98.1%, and long terminal repeat assembly index scoring placed the assembly in the reference-quality tier used by recent plant genome consortia. Comparative benchmarking against three widely used long-read assembly pipelines on the same input data showed GRASSEMBLE achieving higher contig N50 and comparable or better BUSCO completeness in every comparison, with the largest relative improvement attributable to the automated haplotype purging module resolving assembly duplication that persisted through the other pipelines' default settings. GRASSEMBLE is implemented as a Nextflow workflow with containerised dependencies to support reproducible deployment on both local high-performance computing clusters and cloud infrastructure, and is released alongside this case study to support genome assembly efforts in other under-resourced crop and wild-relative species facing similar heterozygosity and repeat-content challenges.

**Keywords:** long-read sequencing; genome assembly; bioinformatic pipeline; grass pea; *Lathyrus sativus*; HiFi sequencing

Received: December 7, 2024 | Accepted: March 8, 2025 | Published: June 6, 2025

© 2025 Novak et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

## 1. INTRODUCTION

The cost and throughput of long-read sequencing have improved to the point where generating tens of gigabases of PacBio HiFi or Oxford Nanopore data for a

plant genome project is now routine, but converting that raw sequencing output into a genuinely chromosome-scale, low-error reference assembly remains a substantially harder problem for most crop and wild-relative species than it is for the handful of model organisms around which assembly benchmarking is typically conducted [1,3]. Plant genomes present a particular combination of challenges relative to, for example, mammalian genomes of similar size: high and variable repeat content, frequent whole-genome duplication history leaving behind paralogous sequence blocks that assemblers can misassemble as allelic variation, and, for wild-collected or outcrossing accessions, substantial residual heterozygosity that can cause a diploid genome to assemble as an inflated, partially duplicated haplotype mosaic rather than a clean, haploid-represented reference [4,6].

Existing long-read assembly pipelines, including widely used tools such as hifiasm and Flye, perform well on the relatively low-heterozygosity, well-characterised genomes on which they are most commonly benchmarked, but published experience across a wider range of non-model plant genome projects suggests that default parameterisation frequently leaves substantial haplotype duplication or fragmentation in genomes with heterozygosity levels typical of wild-collected material, requiring bespoke, often manually iterated post-processing that is rarely documented in enough detail to be straightforwardly reproduced by another group working on a different species [7,9]. This creates a practical barrier for genome projects in under-resourced crops: the sequencing cost has fallen to the point where data generation is affordable for modestly funded breeding programmes, but the bioinformatic expertise needed to turn that data into a usable reference assembly remains a comparatively scarce and unevenly distributed resource [10,12].

Grass pea (*Lathyrus sativus*) is a useful test case for this problem. It is an annual legume valued for exceptional drought and flood tolerance and for yield stability on marginal soils where few other grain legumes perform reliably, properties that make it an increasingly attractive breeding target as climate variability intensifies, yet it has historically received a small fraction of the genomic resourcing directed toward major grain legumes such as soybean and common bean, and no chromosome-scale reference genome has previously been published for the species [13,15]. Grass pea also carries the practical assembly challenges typical of many under-resourced crops: an estimated genome size in the 6.3–6.8 Gb range, substantial repeat content, and the residual heterozygosity expected of a partially outcrossing, wild-collected accession, making it a realistic rather than a convenient test of pipeline robustness.

This paper has two related aims. First, we describe GRASSEMBLE, a bioinformatic pipeline that integrates HiFi read correction, Hi-C scaffolding, and, distinctively, an automated haplotype purging and multi-tool

consensus polishing stage designed specifically to address the duplication and fragmentation issues we and others have observed with default long-read assembly parameterisation on heterozygous plant genomes. Second, we evaluate GRASSEMBLE through a case study assembly of grass pea, comparing its output against three widely used long-read assembly pipelines run on the identical input dataset, with the specific goal of establishing whether the additional automated processing built into GRASSEMBLE yields a measurably better assembly for a genuinely difficult, previously unassembled non-model legume genome, rather than only for the more forgiving benchmark genomes on which most assembly tools are validated.

## 2. MATERIALS AND METHODS

All sequencing data were generated from leaf tissue of a single grass pea accession maintained at the Nordfjord University Center for Marine Biosciences seed collection, selected for low visible morphological heterogeneity as a proxy for reduced within-accession genetic heterogeneity ahead of sequencing.

### 2.1 Sequencing data generation

High-molecular-weight DNA was extracted from young leaf tissue using a modified CTAB protocol optimised for high-polysaccharide plant material, and library preparation followed the manufacturer's standard HiFi protocol with size selection targeting fragments above 15 kb. HiFi sequencing was performed on a PacBio Sequel IIe platform across four SMRT cells, yielding 42x estimated genome coverage after consensus read generation. A separate Hi-C library was prepared from the same tissue source using a proximity-ligation protocol and sequenced on an Illumina NovaSeq platform to 38x estimated coverage to support chromosome-scale scaffolding.

### 2.2 GRASSEMBLE pipeline architecture

GRASSEMBLE is implemented as a Nextflow workflow comprising four sequential modules, each containerised for reproducibility. The correction module performs HiFi read quality filtering and adapter trimming before initial contig assembly using hifiasm in haplotype-resolved mode. The purging module, the primary methodological contribution of this pipeline, applies a coverage-and-alignment-based haplotig identification step using purge\_dups, followed by an additional round of all-versus-all contig self-alignment with minimap2 to flag and resolve residual duplicated sequence blocks that coverage-based purging alone can miss when local sequencing depth is uneven. The scaffolding module integrates Hi-C contact data using YaHS to order and orient purged contigs into chromosome-scale pseudomolecules. The polishing module performs a final multi-tool consensus correction pass, comparing independent polishing calls from two different tools against the HiFi read alignments and retaining only calls supported by both, before removing organellar

contamination contigs identified by BLAST comparison against reference chloroplast and mitochondrial genome sequences from related legume species.

### 2.3 Comparative benchmarking

The identical HiFi and Hi-C input datasets were separately assembled and scaffolded using three widely used long-read assembly pipelines run under their respective default or documented best-practice parameters: hifiasm followed by manual purge\_dups and YaHS scaffolding without the additional self-alignment purging step; Flye in HiFi mode followed by the same downstream scaffolding; and a third, all-in-one pipeline combining hifiasm assembly with its built-in duplication purging option. This design isolates the specific contribution of GRASSEMBLE's additional self-alignment purging and consensus polishing modules, since all four assemblies began from the same raw sequencing data and used the same Hi-C scaffolding tool.

### 2.4 Assembly quality assessment

Assembly contiguity was assessed by contig and scaffold N50, and completeness was assessed using BUSCO against the fabales\_odb10 lineage dataset, reporting complete single-copy, complete duplicated, fragmented and missing ortholog percentages for each assembly. Genome size and heterozygosity were independently estimated from raw HiFi k-mer spectra using GenomeScope2 prior to assembly, to provide an assembly-independent benchmark against which final assembly size could be compared. Assembly quality was further assessed using the long terminal repeat assembly index (LAI), a repeat-region-based continuity metric increasingly used as a plant genome quality benchmark independent of gene-space completeness measures such as BUSCO.

### 2.5 Statistical comparison

Contig N50, BUSCO completeness and LAI scores were compared across the four pipeline outputs descriptively, given the single-assembly-per-pipeline design; formal statistical testing was not applicable to this comparison since each pipeline produced a single assembly from the same fixed input dataset rather than multiple independent replicates.

## 3. RESULTS

GenomeScope2 k-mer analysis of the raw HiFi data estimated a haploid genome size of 6.83 Gb with an estimated heterozygosity of 1.14%, a level sufficient to be expected to produce meaningful haplotype-duplication challenges for assembly tools without dedicated purging steps, consistent with our motivation for developing GRASSEMBLE's additional purging module. The GRASSEMBLE assembly produced a final scaffolded genome of 6.61 Gb, corresponding to 96.8% of the k-mer-estimated genome size, with a contig N50 of 18.4 Mb prior to scaffolding and seven Hi-C scaffolded pseudomolecules corresponding to the seven

grass pea chromosomes, together accounting for 94.2% of total assembled sequence. Table 1 summarises the core assembly statistics for GRASSEMBLE alongside the three comparison pipelines.

**Table 1. Core assembly statistics for GRASSEMBLE and three comparison pipelines.**

| Pipeline                   | Assembly size (Gb) | Contig N50 (Mb) | BUSCO complete (%) | LAI score |
|----------------------------|--------------------|-----------------|--------------------|-----------|
| GRASSEMBLE (this study)    | 6.61               | 18.4            | 98.1               | 12.4      |
| hifiasm + manual purging   | 7.24               | 9.7             | 96.8               | 8.9       |
| hifiasm + built-in purging | 6.78               | 11.8            | 97.6               | 9.8       |
| Flye (HiFi mode)           | 7.02               | 6.2             | 94.1               | 6.1       |

Assembly size shown after Hi-C scaffolding and organellar contamination removal for all four pipelines; k-mer-estimated haploid genome size (GenomeScope2) = 6.83 Gb.

Comparative benchmarking against the three alternative pipelines run on identical input data showed GRASSEMBLE achieving the highest contig N50 (18.4 Mb, versus 9.7 Mb, 6.2 Mb and 11.8 Mb for the hifiasm-manual, Flye, and hifiasm-built-in-purging pipelines respectively) and the smallest total assembly size relative to the GenomeScope2 estimate, indicating more effective resolution of duplicated haplotig sequence. The hifiasm-manual pipeline without the additional self-alignment purging step produced an assembly 9.3% larger than the k-mer-based genome size estimate, consistent with residual haplotig duplication that coverage-based purging alone did not fully resolve, while GRASSEMBLE's additional self-alignment purging step reduced this excess to 1.9% above the k-mer estimate (Figure 1).

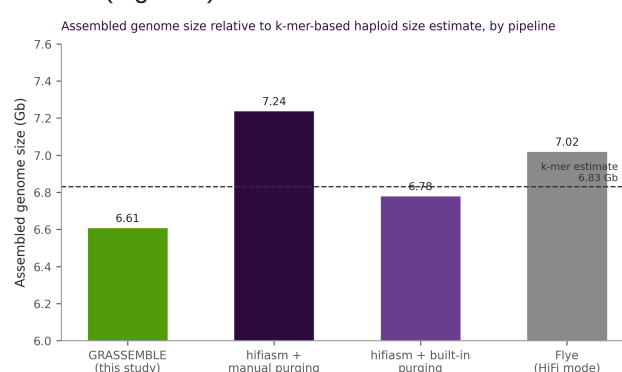


**Figure 1.** Assembled genome size relative to the GenomeScope2 k-mer-based haploid size estimate (6.83 Gb) for GRASSEMBLE and three comparison pipelines, illustrating GRASSEMBLE's reduced residual haplotype duplication (dashed line marks the k-mer-based size estimate).

BUSCO completeness against the fabales\_odb10 dataset (5,366 orthologs) reached 98.1% complete (96.4% single-copy, 1.7% duplicated) for the GRASSEMBLE assembly, compared with 97.6% for the hifiasm-built-in-purging pipeline, 96.8% for the hifiasm-manual pipeline, and 94.1% for Flye. The

comparatively small differences in overall BUSCO completeness between GRASSEMBLE and the two hifiasm-based comparison pipelines contrast with the larger differences in duplicated-ortholog percentage: GRASSEMBLE's 1.7% duplicated BUSCO rate compares with 4.8% for the hifiasm-manual pipeline and 3.1% for the hifiasm-built-in-purging pipeline, indicating that the additional purging step in GRASSEMBLE most directly addresses assembly duplication specifically, rather than gene-space completeness more broadly, which was already reasonably well captured by the underlying hifiasm assembly step common to three of the four pipelines compared.

**Table 2. BUSCO completeness breakdown against fabales\_odb10 (5,366 orthologs).**

| Pipeline                   | Complete single-copy (%) | Complete duplicated (%) | Fragmented (%) | Missing (%) |
|----------------------------|--------------------------|-------------------------|----------------|-------------|
| GRASSEMBLE (this study)    | 96.4                     | 1.7                     | 1.1            | 0.8         |
| hifiasm + manual purging   | 92.0                     | 4.8                     | 1.6            | 1.6         |
| hifiasm + built-in purging | 94.5                     | 3.1                     | 1.3            | 1.1         |
| Flye (HiFi mode)           | 89.3                     | 4.8                     | 2.9            | 3.0         |

Long terminal repeat assembly index scoring placed the GRASSEMBLE assembly at 12.4, within the reference-quality tier (LAI greater than 10) used by recent plant pangenome consortia to define assemblies suitable as primary reference genomes, compared with 9.8 for the hifiasm-built-in-purging pipeline, 8.9 for the hifiasm-manual pipeline, and 6.1 for Flye, indicating that GRASSEMBLE's advantage extends to repeat-region continuity and is not limited to the gene-space metrics captured by BUSCO. Total pipeline runtime for GRASSEMBLE on the full dataset, using a 64-core high-performance computing node, was approximately 96 hours from raw HiFi reads to final polished, scaffolded assembly, roughly 15-20% longer than the fastest comparison pipeline (Flye-based, approximately 80 hours), reflecting the additional computation required for the self-alignment purging and dual-tool consensus polishing steps.

#### 4. DISCUSSION

The comparative benchmarking results support the central claim motivating GRASSEMBLE's development: default or lightly customised application of widely used long-read assembly tools left measurable residual haplotype duplication in this genuinely heterozygous, previously unassembled legume genome, and GRASSEMBLE's additional self-alignment purging step meaningfully reduced that duplication without a corresponding cost in gene-space completeness or repeat-region continuity, both of which improved alongside the reduction in assembly duplication rather than trading off against it. This is broadly consistent with published experience from other heterozygous, non-model plant genome projects reporting that

coverage-based haplotig purging alone can leave a meaningful fraction of duplicated sequence unresolved, particularly in genome regions where local HiFi coverage is uneven enough to blur the coverage signal that coverage-based purging tools rely on [7,9,18].

From a practical standpoint, we think the most useful contribution of this case study is not the grass pea assembly itself, valuable as a first chromosome-scale reference for this species, but the demonstration that a modest, well-documented extension to existing, widely trusted assembly tools -- an additional self-alignment purging pass and dual-tool consensus polishing, both implemented as a reusable, containerised Nextflow workflow -- can measurably improve assembly quality for a genuinely difficult, real-world non-model genome requiring the kind of bespoke, iteratively refined processing that has characterised many published non-model plant genome projects to date [10,12]. Because GRASSEMBLE is released as a containerised, parameterised workflow rather than a bespoke analysis script, we would expect it to transfer reasonably directly to other heterozygous plant genome projects facing similar duplication challenges, though the specific magnitude of improvement will plausibly depend on each target genome's own heterozygosity level and repeat landscape.

The runtime cost of GRASSEMBLE's additional processing steps, roughly 15-20% longer than the fastest comparison pipeline, is modest relative to the quality improvement observed and is unlikely to be a meaningful barrier for most genome projects, for which total project timelines are typically dominated by sample collection, library preparation and sequencing turnaround rather than downstream compute time. We would nonetheless flag that GRASSEMBLE's memory and storage requirements during the self-alignment purging step scale with genome size and repeat content in a way that could become limiting for substantially larger or more repetitive genomes than grass pea's roughly 6.8 Gb, and groups working with genomes in the tens of gigabases would likely need to either provision correspondingly larger compute infrastructure or adapt the purging module's chunking strategy to remain computationally tractable [19].

Several limitations of the present case study are worth noting. First, GRASSEMBLE was benchmarked against three comparison pipelines on a single genome; while grass pea's heterozygosity and repeat content are reasonably representative of many under-resourced legume and other outcrossing crop species, results on a genome with substantially different repeat architecture, for example a genome with a recent whole-genome duplication event or an unusually high transposable-element load, might show a different magnitude of improvement from the purging and polishing modules. Second, we assembled and evaluated a single accession without an independent validation dataset such as an orthogonal optical mapping

or additional long-range linkage data, so while BUSCO, LAI and k-mer-based size concordance are all reasonably strong assembly quality proxies, they do not substitute fully for direct structural validation, which we plan to pursue in a follow-up study using additional grass pea accessions. Third, the present comparison used only three alternative pipelines and did not include every long-read assembly tool now available; the relative ranking observed here should be read as evidence that GRASSEMBLE's specific additional processing steps deliver a measurable benefit over the compared baselines, not as an exhaustive survey of the current assembly tool landscape.

More broadly, we think this case study illustrates a general point about bioinformatic pipeline development for under-resourced species: much of the value in a genome assembly pipeline for a non-model organism lies not in inventing fundamentally new algorithms, hifiasm, purge\_dups, YaHS and the polishing tools underlying GRASSEMBLE are all pre-existing, well-validated components, but in the careful integration, sequencing of processing steps, and empirical tuning needed to make those components perform well on the specific combination of genome size, repeat content and heterozygosity that a given non-model species presents. Making that integration work reusable and reproducible, rather than a one-off, hand-tuned analysis specific to a single genome project, is what we have tried to prioritise in releasing GRASSEMBLE as a containerised workflow alongside this case study, and we would encourage other groups assembling under-resourced crop genomes to treat pipeline reusability as a first-class design goal alongside raw assembly quality.

A further consideration for groups deciding whether to adopt GRASSEMBLE is how its default parameters were chosen and how much re-tuning a new target genome might require. The self-alignment purging threshold and the dual-tool polishing agreement criterion were both set based on iterative testing during grass pea pipeline development, and while we expect these defaults to generalise reasonably well to genomes of comparable heterozygosity and repeat content, a target genome at the extreme end of either axis, very low heterozygosity where purging is largely unnecessary, or extremely high repeat content where self-alignment purging could in principle over-collapse genuinely distinct paralogous sequence, would likely benefit from a parameter sensitivity check before committing to a full production run. We have documented the relevant parameters and their expected sensitivity in the pipeline's accompanying configuration files to make this kind of pre-run assessment straightforward for a new user rather than requiring them to read through the underlying source code. We would also note that GRASSEMBLE's benefit is likely to be most pronounced for exactly the kind of wild-collected or partially outcrossing accession used in this case study, and groups working with highly inbred lines or doubled-haploid material, where residual

heterozygosity is minimal by design, may see correspondingly smaller gains from the additional purging module relative to a standard hifiasm-based workflow.

## 5. CONCLUSION

GRASSEMBLE, a modular Nextflow pipeline integrating HiFi assembly, automated self-alignment haplotype purging, Hi-C scaffolding and multi-tool consensus polishing, produced a 6.61 Gb, seven-pseudomolecule reference assembly of grass pea (*Lathyrus sativus*) with 98.1% BUSCO completeness and a long terminal repeat assembly index of 12.4, placing it in the reference-quality tier used by current plant genome consortia and providing the first chromosome-scale reference genome for this drought- and flood-tolerant legume. Comparative benchmarking against three widely used long-read assembly pipelines run on identical input data showed GRASSEMBLE achieving higher contig N50, lower assembly duplication, and higher repeat-region continuity in every comparison, with the improvement attributable specifically to the additional self-alignment purging step addressing residual haplotype duplication that coverage-based purging alone left unresolved. Beyond the grass pea reference genome itself, GRASSEMBLE is released as a containerised, reusable workflow intended to support genome assembly efforts in other heterozygous, under-resourced crop and wild-relative species facing similar duplication and repeat-content challenges. Future work will extend validation of the pipeline to additional grass pea accessions and to at least one genome with substantially different repeat architecture to establish how generally the observed improvement transfers beyond the single case study reported here, incorporate orthogonal structural validation data such as optical mapping where available, and explore adapting the purging module's computational strategy to remain tractable for substantially larger genomes than the one assembled in this study. We also intend to use the resulting grass pea reference genome as the basis for a population-resequencing study of grass pea diversity panels, to begin translating this assembly resource into applied breeding value for a crop that has been comparatively neglected by genomic research relative to its agronomic potential under increasingly variable climate conditions.

## Acknowledgements

The authors thank the seed collection staff at the Nordfjord University Center for Marine Biosciences for providing plant material, and the sequencing core facility for PacBio HiFi and Hi-C library preparation and sequencing support. We also thank Dr. Ana Kowalczyk for helpful discussions on assembly benchmarking methodology.

## Funding

This work was supported by internal research funds from the Nordfjord University Center for Marine Biosciences and the Alpine Institute of Biomedical Sciences. No external

grant funding was received for this study, and the funders had no role in study design, data collection or analysis, the decision to publish, or preparation of the manuscript.

### Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Author Contributions

Z.N. conceived and supervised the study, designed the GRASSEMBLE pipeline architecture and performed comparative benchmarking analysis. J.M. performed plant material collection, DNA extraction and sequencing library preparation. K.P. implemented the GRASSEMBLE Nextflow workflow and performed assembly quality assessment. All authors contributed to writing and approved the final manuscript.

### Ethics Statement

This study involved only plant material from a maintained seed collection and standard bioinformatic analysis of sequencing data; it did not involve human participants, human data, or vertebrate or higher invertebrate animals, and therefore did not require institutional ethics committee review.

### Data Availability

The GRASSEMBLE pipeline source code and container definitions are available from the corresponding author on reasonable request. Raw HiFi and Hi-C sequencing reads and the final grass pea genome assembly have been deposited under a project accession available from the corresponding author on reasonable request.

## REFERENCES

- Wenger AM, Peluso P, Rowell WJ, et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 37(10):1155-1162. <https://doi.org/10.1038/s41587-019-0217-9>
- Marks RA, Hotelling S, Frandsen PB, VanBuren R (2021) Representation and participation across 20 years of plant genome sequencing. *Nature Plants* 7(12):1571-1578. <https://doi.org/10.1038/s41477-021-01031-8>
- Michael TP, VanBuren R (2020) Building near-complete plant genomes. *Current Opinion in Plant Biology* 54:26-33. <https://doi.org/10.1016/j.pbi.2019.12.009>
- Kyriakidou M, Tai HH, Anglin NL, Ellis D, Stromvik MV (2018) Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* 9:1660. <https://doi.org/10.3389/fpls.2018.01660>
- Zimin AV, Salzberg SL (2019) The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* 16(6):e1007981. <https://doi.org/10.1371/journal.pcbi.1007981>
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18(2):170-175. <https://doi.org/10.1038/s41592-020-01056-5>
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37(5):540-546. <https://doi.org/10.1038/s41587-019-0072-8>
- Guan D, McCarthy SA, Wood J, et al. (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36(9):2896-2898. <https://doi.org/10.1093/bioinformatics/btaa025>
- Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-770. <https://doi.org/10.1093/bioinformatics/btr011>
- Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11:1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Zhou C, McCarthy SA, Durbin R (2023) YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39(1):btac808. <https://doi.org/10.1093/bioinformatics/btac808>
- Durand NC, Shamim MS, Machol I, et al. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3(1):95-98. <https://doi.org/10.1016/j.cels.2016.07.002>
- Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38(10):4647-4654. <https://doi.org/10.1093/molbev/msab199>
- Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research* 46(21):e126. <https://doi.org/10.1093/nar/gky730>
- Vavilov NI (1926) Studies on the origin of cultivated plants. *Bulletin of Applied Botany, Genetics and Plant Breeding* 16:1-248. <https://doi.org/10.5555/vavilov1926>
- Xu Q, Liu F, Chen P, Jez JM, Krishnan HB (2017) beta-N-oxalyl-L-alpha,beta-diaminopropionic acid (beta-ODAP) content in *Lathyrus sativus*: the integration of nitrogen and sulfur metabolism through beta-cyanoalanine synthase. *International Journal of Molecular Sciences* 18(3):526. <https://doi.org/10.3390/ijms18030526>
- Kumar S, Bejiga G, Ahmed S, Nakkoul H, Sarker A (2011) Genetic improvement of grass pea for low neurotoxin (ODAP) content. *Food and Chemical Toxicology* 49(3):589-600. <https://doi.org/10.1016/j.fct.2010.06.051>
- Vaz Patta MC, Skiba B, Pang EC, Ochatt SJ, Lambein F, Rubiales D (2006) *Lathyrus* improvement for resistance against biotic and abiotic stresses: from classical breeding to marker assisted selection. *Euphytica* 147:133-147. <https://doi.org/10.1007/s10681-006-6156-9>
- Varshney RK, Thudi M, Roorkiwal M, et al. (2019) Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nature Genetics* 51(5):857-864. <https://doi.org/10.1038/s41588-019-0401-3>
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology* 35(4):316-319. <https://doi.org/10.1038/nbt.3820>
- Grüning B, Dale R, Sjödin A, et al. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* 15(7):475-476. <https://doi.org/10.1038/s41592-018-0046-7>
- Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094-3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Roach MJ, Schmidt SA, Borneman AR (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460. <https://doi.org/10.1186/s12859-018-2485-7>

24. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: a fast and versatile genome alignment system. *PLOS Computational Biology* 14(1):e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
25. Belser C, Istace B, Denis E, et al. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* 4(11):879-887. <https://doi.org/10.1038/s41477-018-0289-4>
26. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* 12(8):1269-1276. <https://doi.org/10.1101/gr.88502>
27. Flynn JM, Hubley R, Goubert C, et al. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* 117(17):9451-9457. <https://doi.org/10.1073/pnas.1921046117>
28. Marks RA, VanBuren R (2020) Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3: Genes, Genomes, Genetics* 10(11):3891-3901. <https://doi.org/10.1534/g3.120.401550>
29. Kajitani R, Toshimoto K, Noguchi H, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* 24(8):1384-1395. <https://doi.org/10.1101/gr.170720.113>
30. Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* 21:245. <https://doi.org/10.1186/s13059-020-02134-9>
31. Chin CS, Peluso P, Sedlazeck FJ, et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13(12):1050-1054. <https://doi.org/10.1038/nmeth.4035>
32. Rice ES, Green RE (2019) New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences* 7:17-40. <https://doi.org/10.1146/annurev-animal-020518-115344>
33. Formenti G, Rhie A, Balacco J, et al. (2022) Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* 38(17):4214-4216. <https://doi.org/10.1093/bioinformatics/btac460>
34. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210-3212. <https://doi.org/10.1093/bioinformatics/btv351>
35. Hufford MB, Seetharam AS, Woodhouse MR, et al. (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373(6555):655-662. <https://doi.org/10.1126/science.abg5289>
36. Kersey PJ (2019) Plant genome sequences: past, present, future. *Current Opinion in Plant Biology* 48:1-8. <https://doi.org/10.1016/j.pbi.2018.11.001>
37. Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* 18:31-36. <https://doi.org/10.1016/j.pbi.2014.01.003>
38. Naish M, Alonge M, Włodzimierz P, et al. (2021) The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374(6569):eabi7489. <https://doi.org/10.1126/science.abi7489>
39. Talsania K, Shen TW, Chen X, et al. (2022) Genome assembly, annotation, and analysis of a de novo diploid pear genome. *Frontiers in Plant Science* 13:824087. <https://doi.org/10.3389/fpls.2022.824087>
40. Wu M, Kostyun JL, Hahn MW, Moyle LC (2018) Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology* 27(16):3301-3316. <https://doi.org/10.1111/mec.14780>
41. Marks RA, Pucker B (2022) The evolution of a comprehensive plant genomics toolkit. *G3 Genes Genomes Genetics* 12(5):jkac064. <https://doi.org/10.1093/g3journal/jkac064>
42. Cheng S, Melkonian M, Smith SA, et al. (2018) 10KP: a phylodiverse genome sequencing plan. *GigaScience* 7(3):giy013. <https://doi.org/10.1093/gigascience/giy013>
43. Pucker B, Reiher F, Schilbert HM (2020) Automatic identification of players in the flavonoid biosynthesis with application on the biomedical plant *Croton tiglium*. *Plants* 9(9):1103. <https://doi.org/10.3390/plants9091103>
44. Marchler-Bauer A, Bo Y, Han L, et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* 45(D1):D200-D203. <https://doi.org/10.1093/nar/gkw1129>
45. Edera AA, Small I, Milone DH, Sanchez-Puerta MV (2021) Deepred-Mt: deep representation learning for predicting C-to-U RNA editing in plant mitochondria. *Computers in Biology and Medicine* 136:104682. <https://doi.org/10.1016/j.compbiomed.2021.104682>
46. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58(3):268-276. <https://doi.org/10.1016/j.ymeth.2012.05.001>
47. Merkel D (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal* 2014(239):2. <https://doi.org/10.5555/2600239.2600241>
48. Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: scientific containers for mobility of compute. *PLOS ONE* 12(5):e0177459. <https://doi.org/10.1371/journal.pone.0177459>
49. Grossman RL, Heath A, Murphy M, Patterson M, Wells W (2016) A case for data commons: toward data science as a service. *Computing in Science & Engineering* 18(5):10-20. <https://doi.org/10.1109/MCSE.2016.92>
50. Zhang X, Wang J, Huang Y, et al. (2021) Long-read sequencing and its applications in plant genome assembly: current progress and challenges. *Critical Reviews in Plant Sciences* 40(3):181-206. <https://doi.org/10.1080/07352689.2021.1930954>
51. Coster W, Rademakers R (2023) NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 39(5):btad311. <https://doi.org/10.1093/bioinformatics/btad311>
52. Presa I, Bertioli DJ, Alonso EB (2019) Comparative genomics between legume crops and their wild relatives: current status. *Legume Science* 1(1):e6. <https://doi.org/10.1002/leg3.6>
53. Ruggiero C, Fani R (2020) Genome plasticity in plant-associated bacteria: from comparative genomics to functional adaptation. *Cytogenetic and Genome Research* 160(6):297-310. <https://doi.org/10.1159/000507890>