Reference genome assembly of the tetraploid tuber crop *Tropaeolum tuberosum*

Ina Scheffler, Thomas Greb and Pascal Hunziker

Centre for Organismal Studies (COS), Heidelberg University, Germany

Cluster of Excellence GreenRobust, Heidelberg University, Germany

Corresponding author contact information:

Pascal Hunziker, Centre for Organismal Studies (COS), Heidelberg University, Im Neuenheimer Feld 360, 69120 Heidelberg, Germany

phone: +49-(0)6221-54-5785

email: pascal.hunziker@cos.uni-heidelberg.de

Abstract

Tropaeolum tuberosum is a tetraploid tuber-forming crop with agroecological and agronomic potential, yet genomic resources for this species remain scarce and limit genetic and functional studies. To address this gap, we generated a reference genome assembly for *T. tuberosum* using PacBio HiFi reads with an estimated genome size of 418 Mb based on *k*-mer analysis. The final assembly spans 1.32 Gb with 2,189 contigs (contigs N50 = 32.2 Mb, longest contig = 60 Mb) and recovers 79 % of the estimated genome size. We assessed assembly completeness and accuracy using Benchmarking Universal Single-Copy Orthologs (BUSCO), which detected 98.8 % complete genes (21.6 % single-copy, 77.1 % duplicated), 0.5 % fragmented, and 0.7 % missing, demonstrating near-complete gene space recovery consistent with a high-quality tetraploid reference genome. Repetitive sequences account for 71.6 % of the genome, and we annotated 87,927 protein-coding genes using Helixer. This reference genome assembly represents the first genome-scale resource for *T. tuberosum* and will enable studies of evolution, domestication and comparative genomics, and support breeding, conservation, and functional genomics in this species and related taxa.

Introduction

Tropaeolum tuberosum Ruiz & Pav. (2n = 4x = 52) is a tuber-forming crop traditionally cultivated in the high plains of the Andean Altiplano at elevations exceeding 3,500 m above sea level. The species is adapted to challenging environmental conditions, including strong diurnal temperature fluctuations and nutrient-poor soils, but produces tubers with high nutritional value [1]. Its tubers are rich in carbohydrates, proteins, vitamins, and a diverse array of secondary metabolites, including high levels of glucosinolates with antioxidant and antimicrobial activities [2, 3]. These characteristics make *T. tuberosum* a promising candidate for resilient and sustainable agricultural systems. Although glucosinolates are defining metabolites of the *Brassicales*, their biochemical diversity and evolutionary distribution outside the Brassicaceae remain incompletely characterized [4], underscoring the value of genomic resources from non-brassicaceous lineages.

T. tuberosum belongs to the genus *Tropaeolum* (family *Tropaeolaceae*, order *Brassicales*), a Neotropical lineage of roughly 80 herbaceous species that includes ornamental nasturtiums and several tuber-forming taxa adapted to cool, often high-elevation environments [5]. Despite its agronomic and nutritional potential, *T. tuberosum* remains an underutilized crop [3]. In contrast to other tuber crops, such as potato, it has not been subject of extensive breeding

programs, and its genetic improvement has largely relied on traditional cultivation and farmer selection. As a result, genomic and molecular resources for *T. tuberosum* are scarce.

T. tuberosum is a tetraploid species (2n = 52), with cytological evidence of multivalents and meiotic irregularities suggesting autopolyploid origin, though segmental allopolyploidy involving hybridization with closely related taxa such as *T. cochabambe* remains under discussion [6]. Polyploidy can contribute to phenotypic robustness and environmental tolerance, but also complicates genome assembly, gene annotation and the interpretation of gene dosage and expression patterns [7]. High-quality genome assemblies are therefore particularly important for polyploid crops, as they provide a framework for studying genome organization, regulatory interactions, and the consequences of whole-genome duplication.

To support genomic research and future improvement efforts, we present here a long-read-based genome assembly of tetraploid *T. tuberosum* using accession BG HEID 007454. This accession has been maintained *ex situ* in European botanical gardens for at least three decades under vegetative propagation, providing genetically uniform material advantageous for polyploid genome assembly where heterozygosity and somatic variation can complicate haplotype resolution. Long-term cultivation outside the Andean Altiplano — with altered photoperiod, temperature dynamics, and radiation regimes — has resulted in an accession that produces tubers reliably under temperate autumn photoperiods, suggesting adaptation or selection for cultivation outside its native range. This offers a unique perspective for investigating polyploid genome stability and the genetic basis of environmental responsiveness in tuber crops [8, 9].

In contrast to model *Brassicales* lineages, genomic resources for *Tropaeolum* and *T. tuberosum* remain sparse. While a recent preprint reported a genome assembly for *Tropaeolum majus* [10], no reference genome has been available for *T. tuberosum*. Extending haplotype-resolved assembly approaches to *Tropaeolum* fills a major gap in *Brassicales* comparative genomics and provides the first comprehensive genomic framework for investigating glucosinolate metabolism, tuber development, and environmental adaptation in a non-brassicaceous lineage within the *Brassicales*. Moreover, the genome assembly presented in this work establishes a foundational genomic resource for *T. tuberosum*. It enables analyses of polyploid genome organization, supports comparative studies with other tuber crops, and provides a framework for investigating traits related to tuberization, environmental responses, and specialized metabolism. Ongoing work is leveraging this resource for population-scale diversity analysis, synteny mapping and functional genetics. More broadly, this resource contributes to ongoing efforts to develop genomic tools for

neglected and underutilized crops and supports research aimed at expanding their potential across diverse agricultural contexts.

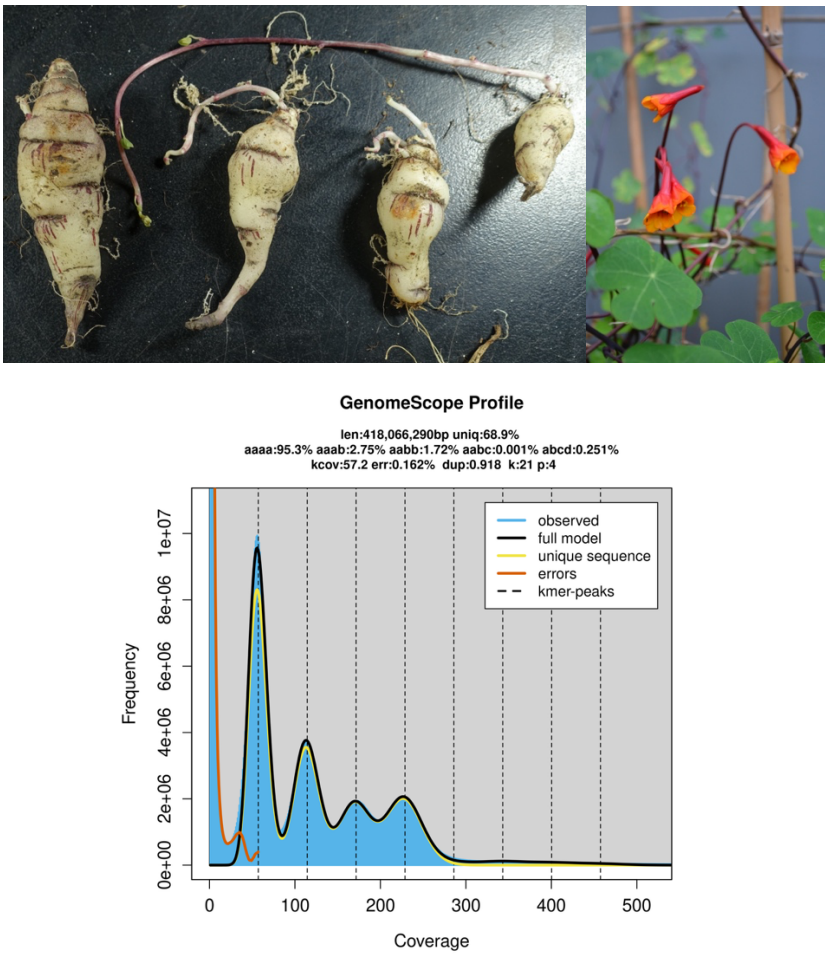

Figure 1: Genome of *T. tuberosum*. Tubers, leaves and flowers of the BG HEID 007454 accession, and distribution profile of 21-mer analysis.

Results & Discussion

We sequenced and assembled the genome of *Tropaeolum tuberosum* Ruiz & Pav (Tropaeolaceae*, Brassicales*) accession BG HEID 007454 from the living collection of the Botanic Garden Heidelberg, Germany (**Fig. 1**). Using single-molecule real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio), a total of 128.2 Gb of PacBio HiFi sequencing data were produced. *K*-mer analysis revealed an estimated genome size of 418.1 Mb, a heterozygosity rate of 4.7 % and confirmed an autotetraploid genome topology (**Fig. 1**). This size was significantly smaller than the previously reported 1.239 Gb measured using flow cytometry [11]. Ploidity topology of *T. tuberosum* is part of an ongoing discussion, and both

autotetraploid and allotetraploid origins have been postulated [6]. Our *k*-mer analysis showed heterozygosity forms of aaab at 2.75 %, which is larger than the aabb form (1.72 %), indicating that *T. tuberosum* is of autotetraploid origin (**Fig. 1**).

Table 1: Assembly statistics for the *T. tuberosum* genome v1.0

Number of contigs	2189
Total assembly size	1.32 Gbp
Largest contig	60 Mbp
N50 contig length	32.2 Mbp
L50 contig count	16
N90 contig length	1.1 Mbp
L90 contig count	96
GC content	36.73 %
BUSCO v5 embryophyta_odb10	C:98.8% [S:21.6%, D:77.1%], F:0.5%, M:0.7%, n:1614, E:1.9%

Next, we assembled a reference genome using hifiasm (**Table 1**). The assembly contained 2,189 contigs, with a contig N50 of 32.2 Mb and 90 % of the assembled genome contained in 96 contigs. The total assembly size of 1.32 Gb aligned well with the expected ~1.67 Gb tetraploid *T. tuberosum* genome size minus minor haplotype divergence effects. Bandage plots showed larger contigs having around 53x depth, as expected, while numerous small contigs with 1-3x depth are displayed that could stem from contamination, organellar DNA fragments, sequencing artifacts or extremely divergent haplotype regions (**Suppl. Fig. 1**).

Our assembly showed 98.8 % complete Benchmarking Universal Single-Copy Orthologs (BUSCOs), with a high proportion of duplicated BUSCOs (77.1 %) and minimal fragmentation (0.5 %), reflecting the tetraploid nature of the genome and the separation of homoeologous loci into haplotype-resolved contigs (**Table 1**). Merqury *k*-mer analysis of the tetraploid genome showed that individual haplotype assemblies captured ~66–71 % of read *k*-mers each, whereas combining haplotypes recovered 95.1 % of the total *k*-mer space. Consensus quality (QV) range was 59–65 per haplotype, confirming very high sequence accuracy. Together, these results are consistent with a haplotype-resolved, partially phased assembly of a heterozygous tetraploid genome.

De novo repeat annotation using RepeatModeler identified 1,706 repeat families, including 732 LTR retrotransposon families detected using structural criteria (**Table 2**). Repetitive

elements accounted for 71.6 % of the genome, with LTR retrotransposons representing the major fraction (33.6 %), dominated by Gypsy (15.6 %) and Copia (12.9 %) elements. DNA transposons contributed 1.9 %, while 18.7 % of repeats remained unclassified. The overall repeat content of *T. tuberosum* falls within the range reported for other eudicot plant genomes of comparable size [12]. The predominance of LTR retrotransposons, particularly Gypsy and Copia elements, is consistent with patterns observed in many angiosperms and likely contributes substantially to genome size expansion [13]. The proportion of unclassified repeats may reflect lineage-specific elements or limitations of current repeat databases. After removal of redundant models, the final library was used for genome-wide repeat masking and annotation.

Table 2: Repeat landscape of the *T. tuberosum* genome v1.0

	Number of elements	Percentage of sequence
Retroelements	**613664**	**36.85 %**
LINEs:	54466	3.28 %
RTE/Bov-B	296	0 %
L1/CIN4	54170	3.27 %
LTR elements:	559198	33.58 %
Ty1/Copia	234974	12.87 %
Gypsy/DIRS1	201459	15.64 %
DNA transposons	**49786**	**1.91 %**
Hobo-Activator	4185	0.22 %
MULE-MuDR	40307	1.57 %
Tourist/Harbinger	1902	0.07 %
Rolling-circles	**3048**	**0.13 %**
Unclassified:	**680448**	**18.67 %**
Small RNA:	**31591**	**13.2 %**
Simple repeats:	**188502**	**0.7 %**
Low complexity:	**39269**	**0.15 %**

Helixer predicted 87,927 protein-coding genes in the *T. tuberosum* genome, with an average of 4.7 exons per gene and a mean protein length of 364.5 amino acids (**Table 3**). All predicted proteins were complete, with no internal stop codons detected, indicating high-quality gene models. BUSCO analysis revealed exceptional annotation completeness, with 98.1 % of conserved plant genes (1,583 of 1,614 BUSCOs) identified as complete. Of these, 23.2 % were single-copy and 74.8 % were duplicated, consistent with the retention of duplicated

genes following ancient whole genome duplication events common in flowering plants [14]. Only 1.2 % of BUSCOs were fragmented and 0.7 % were missing, demonstrating that our annotation captured nearly the complete gene complement of this species. For comparison, a homology-based GeMoMa annotation of the closely related diploid *Tropaeolum majus* identified 32,924 protein-coding genes with 96.8% BUSCO completeness [10]; however, differences in ploidy and annotation strategy likely contribute to this discrepancy, so we did not attempt a detailed cross-species interpretation.

The protein length distribution showed a typical pattern for plant genomes, with the majority of proteins ranging from 100-500 amino acids (54,092 proteins, 61.5 %). A small fraction of very short proteins (<50 aa; 3,714 proteins, 4.2 %) likely represent peptides or annotation artifacts, while unusually long proteins (>2,000 aa; 816 proteins, 0.9 %) may include multi-domain proteins or potential gene fusions warranting further investigation.

Table 3: Helixer annotation statistics summary for structural annotation of the *T. tuberosum* genome v1.0

Total protein-coding genes	87,927
Mean protein length	364.4 aa
Mean exons per gene	4.7
Proteins without internal stops	100 %
BUSCO v5 embryophyta_odb10	C:98.1 % [S:23.2 %, D:74.8 %], F:1.2 %, M:0.7 %, n:1614
Proteins 100-500 aa	61.5 %
Very short proteins (<50 aa)	4.2 %
Very long proteins (>2000 aa)	0.9 %

This first reference genome assembly for *T. tuberosum* achieves high completeness and sequence quality, providing a foundation for comparative genomics, functional analyses, and diversity studies. The assembly establishes a framework for future refinement through chromosome-scale scaffolding and serves as an anchor for population-scale resequencing efforts. This resource enables investigation of polyploid genome organization, tuber development, and specialized metabolism in *T. tuberosum* and comparative studies across the Tropaeolaceae family.

Methods

Plant growth

20 cm long shoot tip segments were cut from a vegetative *T. tuberosum* plant, leaves removed with scissors, and sterilized by sequentially soaking 5 min in 70 % ethanol and 15 min in 1.5 % sodium hypochlorid, 0.1 % Tween-20, followed by six washes with sterile water. Nodal segments of 1-2 cm length were cut and placed onto 1X MS medium (Musharige and Skoog medium including vitamins (Duchefa M0222), 0.5 g/L MES, 30 g/L sucrose, 8 g/L agar, pH 5.7) in Magenta GA-7 vessels. Plants were grown in a climate chamber at 25°C under long-day conditions (16 h light). 3-week-old plants were used for DNA extraction.

DNA extraction & PacBio sequencing

We extracted high molecular weight DNA from *T. tuberosum* accession BG HEID 007454 maintained in the living collection of the Botanic Garden Heidelberg, Germany. The accession was received from the Botanic Garden Dresden, Germany, in 2025, which had received it from the Botanic Garden Bonn, Germany, in 1995. The accession has been maintained under vegetative propagation in European botanical collections for at least three decades. Historical records prior to 1995 and original geographic provenance are unavailable. High molecular weight DNA was extracted from a single in vitro grown plant using a CTAB-based protocol [15]. DNA quality and quantity were assessed using a NanoDrop spectrophotometer (ND-2200, Thermo Fisher Scientific), Qubit dsDNA BR Assay (Thermo Fisher Scientific), and TapeStation Genomic DNA ScreenTape Assay (Agilent Technologies). DNA was shipped to Novogene Corporation (https://www.novogene.com/). DNA quality was assessed a second time by Novogene Corporation using NanoDrop, Qubit, gel electrophoresis and pulsed-field gel electrophoresis. Library preparation was performed using the SMRTbell Express Template Prep Kit 2.0. and the resulting library was sequenced using one SMRT Cell 25M Tray on a PacBio Revio system.

Genome size estimation

Jellyfish v2.2.10 was used to perform *k*-mer = 21 frequency distribution analysis [16]. Subsequently, GenomeScope v2.0 was used to estimate genome size, heterozygosity, and duplication rate using k = 21 and n = 4 [17].

Genome sequence assembly, repeat modeling and masking

The genome was assembled using hifiasm v0.25.0-r726 [18–20] with hifiasm -o asm -t 48 --hg-size 0.42g. Assembly statistics were generated using gfastats v1.3.11 [21]. Genome quality was assessed using BUSCO v5.8.3 [22] with the embryophyte_odb10 database in genome mode on the primary hifiasm assembly. *K*-mer-based evaluation of genome completeness and accuracy was performed using meryl v1.4.1 and merqury v1.3 [23]. Meryl was used to count canonical *k*-mers (k = 21) from the raw HiFi reads and from the assembled haplotype sequences. Merqury then compared the read k-mer spectra with each haplotype assembly to estimate the proportion of read k-mers captured, the completeness of each haplotype, and base-level consensus quality (QV).

Genome annotation

The primary hifiasm assembly was utilized for *de novo* repeat modeling using RepeatModeler v2.0.7 [24] using rmblast 2.14.1+. Assembly was soft-masked using RepeatMasker v4.2.2 [25]. Gene prediction was performed using Helixer v0.3.2 [26], a deep learning-based gene prediction tool trained on diverse plant genomes. The repeat-masked genome assembly was processed using the pre-trained 'land_plant' model with default parameters optimized for plant gene structures. Helixer was run without overlap mode using a batch size of 4 on an NVIDIA A100 GPU with CUDA 11.8 and cuDNN 8.9.7. Gene models were output in GFF3 format, and protein sequences were extracted using gffread v0.12.7[27]. The completeness and quality of the predicted gene set were assessed using BUSCO v5.4.7 [22] with the embryophyta_odb10 database (n=1,614 conserved plant genes) in protein mode.

Author contributions:

Study conception and design: PH, Plant sample preparation: IS, Experimental work, data analysis and manuscript draft and preparation: PH, Infrastructure and manuscript preparation: TG.

Data Availability:

Raw reads for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB108055 (https://www.ebi.ac.uk/ena/browser/view/PRJEB108055).

Reference genome assembly (v1.0) and annotation files have been deposited at FigShare (https://doi.org/10.6084/m9.figshare.31338463).

Ethics and Compliance:

Plant material (*Tropaeolum tuberosum* accession BG HEID 007454) was obtained under a Material Transfer Agreement from the Botanic Garden Heidelberg, Germany. Maintained in European collections since at least 1995 (predating the Nagoya Protocol and EU Regulation 511/2014), it lacks documented collecting event or provenance. No new access from other countries was required, placing this study outside ABS scope; all institutional conditions were respected.

Acknowledgements:

We thank the Botanical Garden Heidelberg for providing plant material (BG HEID 007454). Thanks to Prof. Dr. Sandra Schmöckel and Dr. David Jarvis for helpful discussions. Thanks to Prof. Dr. Marcus Koch for commenting on the manuscript. The authors acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1803-1 FUGG and INST 35/1804-1 LAGG. The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) for funding under "Project number 455787709" (bwForCluster

BinAC 2). This study has been made possible by funding of the Klaus Tschira Boost Fund, a joint initiative of GSO – Guidance, Skills & Opportunities e.V. and Klaus Tschira Stiftung.

Supplementary Figures

Supplementary Figure 1: Bandage plot colored according to depth using default settings. Yellow color shows nodes with high depth above the 3rd quartile. Median depth is 53x.

References

1. Coloma A, Flores-Mamani E, Quille-Calizaya G, Zaira-Churata A, Apaza-Ticona J, Calsina-Ponce WC, et al. Characterization of Nutritional and Bioactive Compound in Three Genotypes of Mashua (Tropaeolum tuberosum Ruiz and Pavón) from Different Agroecological Areas in Puno. International Journal of Food Science. 2022;2022:7550987.

2. Apaza Ticona LN, Tena Pérez V, Bermejo Benito P. Local/traditional uses, secondary metabolites and biological activities of Mashua (Tropaeolum tuberosum Ruíz & Pavón). J Ethnopharmacol. 2020;247:112152.

3. Luera-Quiñones S, Correa MJ, Moreno-Rojo C, Salvador-Reyes R, Paucar-Menacho LM. Mashua (Tropaeolum tuberosum Ruiz & Pavón): Nutritional Composition, Bioactive Compounds, and Functional Potential as an Andean Natural Ingredient. Foods. 2025;14.

4. Agerbirk N, Koch MA. Inconclusive evidence for chain-elongated methionine-derived glucosinolates beyond the family Brassicaceae. Biochem Syst Ecol. 2025;122:105033.

5. Andersson L, Andersson S. A molecular phylogeny of *Tropaeolaceae* and its systematic implications. Taxon. 2000;49:721–36.

6. Johns T, Neil Towers GH. Isothiocyanates and thioureas in enzyme hydrolysates of Tropaeolum tuberosum. Phytochemistry. 1981;20:2687–9.

7. Shi X, Zhang C, Ko DK, Chen ZJ. Genome-Wide Dosage-Dependent and -Independent Regulation Contributes to Gene Expression and Evolutionary Novelty in Plant Polyploids. Mol Biol Evol. 2015;32:2351–66.

8. Cortinovis G, Di Vittori V, Bellucci E, Bitocchi E, Papa R. Adaptation to novel environments during crop diversification. Curr Opin Plant Biol. 2020;56:203–17.

9. Gutaker RM, Purugganan MD. Adaptation and the geographic spread of crop species. Annu Rev Plant Biol. 2024;75:679–706.

10. Friedhoff R, Wittstock U, Pucker B. Genome sequence of the medicinal plant *Tropaeolum majus* provides insights into flavonoid biosynthesis. BioRxiv. 2024.

11. Veselý P, Bureš P, Šmarda P. Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives. Ann Bot. 2013;112:1193–200.

12. Chanderbali AS, Jin L, Xu Q, Zhang Y, Zhang J, Jian S, et al. Buxus and Tetracentron genomes help resolve eudicot genome history. Nat Commun. 2022;13:643.

13. Zhang S-J, Liu L, Yang R, Wang X. Genome size evolution mediated by gypsy retrotransposons in brassicaceae. Genomics Proteomics Bioinformatics. 2020;18:321–32.

14. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18:411–24.

15. Vaillancourt B, Buell CR. High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. BioRxiv. 2019.

16. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.

17. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432.

18. Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat Methods. 2024;21:967–70.

19. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5.

20. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.

21. Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, et al. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. Bioinformatics. 2022;38:4214–6.

22. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 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.

23. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.

24. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020;117:9451–7.

25. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;Chapter 4:Unit 4.10.

26. Holst F, Bolger AM, Kindel F, Günther C, Maß J, Triesch S, et al. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. Nat Methods. 2025.

27. Pertea G, Pertea M. GFF utilities: gffread and gffcompare. F1000Res. 2020;9.