scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-scale genome assembly of *Tigridiopalma magnifica*

Duc Quy Vu[1,2,3,4,5], Tian-Wen Xiao[1,2,3,4], Zheng-Feng Wang[2,3,4,6] ✉ & Xue-Jun Ge[1,2,3,4] ✉

***Tigridiopalma magnifica*, a member of the Melastomataceae family, is a monotypic genus perennial plant species endemic to South China. Due to its rarity and restricted distribution, the species has been designated as a national second-level protected plant of China. To enhance conservation efforts for *T. magnifica*, a chromosome scale genome was assembled using combined Oxford Nanopore long-read and Hi-C data. The assembled *T. magnifica* genome spans 217,339,795 bp across 22 pseudochromosomes with a scaffold N50 of 11,008,659 bp. Telomeric caps at both ends of all the pseudochromosomes were identified, and no intra-chromosomal gaps existed except of pseudochromosomes 5 and 15 with each containing one. The genome demonstrates 95% Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness, a Long Terminal Repeat Assembly Index (LAI) value of 16.01 and Merqury quality value (QV) of 42.09. Genome annotation revealed 43,291 protein-coding genes and 475 tRNA genes. This high-quality assembly and annotation provide essential insights into the genomic features of *T. magnifica*, facilitating effective conservation strategies and future applications for this species.**

Background & Summary

Many plant species face extinction risks due to evolutionary dead ends and anthropogenic factors (including habitat destruction, pollution, invasive species, over-exploitation, and climate change)[1]. The integration of genomic resources into conservation efforts provides essential information for endangered plants and contributes significantly to biodiversity preservation[2,3]. *Tigridiopalma magnifica* (Melastomataceae), endemic to Southern China, was initially described in 1979[4] as the sole representative of its genus. The species exhibits ornamental potential due to its attractive leaf and flower morphology, particularly its shade tolerance characteristics (Fig. 1)[4]. The distribution of *T. magnifica* is primarily determined by several environmental factors, including canopy closure, soil moisture content, and stream proximity[5]. These specific habitat requirements have contributed to the species endangered status in the wild, leading to its classification as a national second-level key protected wild plant of China[6].

Despite containing approximately 5858 species in Melastomataceae[7], currently, nuclear genomes with annotated proteomes have only been published in two species: *Melastoma candidum* and *Barthea barthei*[8,9], while many others have their genomes at plastome level[10,11] which included the chloroplast genome of *T. magnifica*[12]. With the development of genomic sequencing technologies, the frontier of plant genomics has shifted toward "Telomere-to-Telomere" (T2T) assemblies, which aim to provide gapless, end-to-end chromosomal sequences. These T2T genome assembling generally applies third generation long sequencing reads from platforms of Nanopore or PacBio and Hi-C reads to achieve[13]. With the T2T assembling, highly repetitive areas like centromeres, telomeres, and segmental duplications in the genome can be resolved. More novel and missing genes will be discovered, and the accuracy of (structural) variances detection is then improved[13].

[1]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [2]South China National Botanical Garden, Guangzhou, 510650, China. [3]Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [4]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [5]University of Chinese Academy of Sciences, Beijing, 100049, China. [6]Key Laboratory of Vegetation Restoration and Management of Degraded ecosystems, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. ✉e-mail: wzf@scbg.ac.cn; xjge@scbg.ac.cn

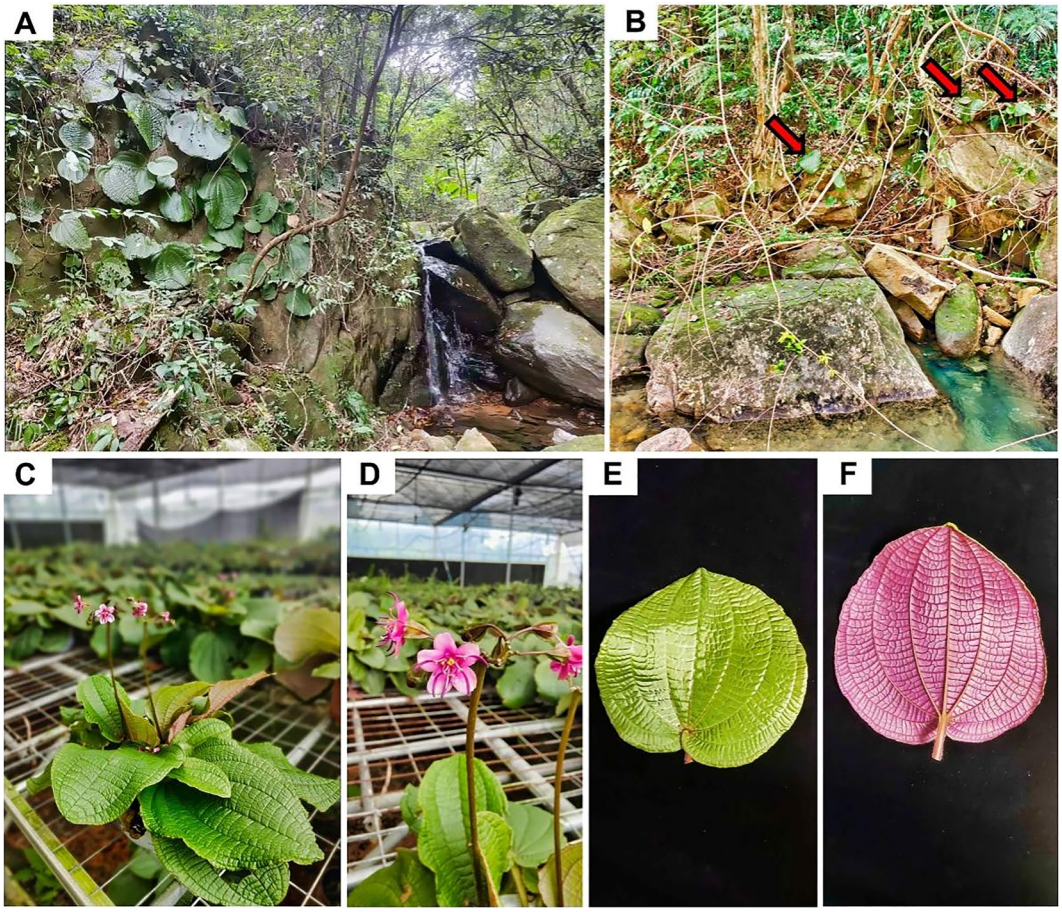

Fig. 1 Picture of *Tigridiopalma magnifica* habitat (**A** and **B**), individual used for sequencing (**C** and **D**), flowers (**D**) and leaves (**E** and **F**).

In Melastomataceae, *T. magnifica* lacks comprehensive genomic resources, especially at T2T level, limiting evidence-based conservation strategies. Here, we present a telomere-to-telomere chromosome-scale and fully annotated genome of *T. magnifica*. The assembly and annotation utilized approximately 139 Gb of long reads from Oxford Nanopore Technology (ONT), 114 Gb of short reads, 28 Gb of high-throughput chromosome conformation capture (Hi-C) reads, and 70 Gb of RNA sequencing (RNA-seq) reads. The initial *T. magnifica* genome assembly generated by Nextdenovo[14] measured 257,764,685 bp with an N50 of 86,41,021 bp. Following Hi-C reads scaffolding by HapHic[15] and manual curation, the assembly measured 217,339,795 bp with an N50 of 11,008,659 bp, anchored to 22 pseudochromosomes. The analysis identified 43,291 genes and 475 tRNAs. The BUSCO[16] evaluation demonstrated 95.0% genome completeness. Additionally, genome quality assessment through LAI[17] indicated a value of 16.01, and quality value (QV) through Merqury[18] of 42.09. This high-quality *T. magnifica* genome assembly provides essential resources for its evolutionary and conservation studies.

Methods

Sample collection and sequencing. Leaf and flower tissues of *T. magnifica* were collected from an individual cultivated at the South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. This individual was propagated through tissue culture methods[19]. The tissues were flash-frozen in liquid nitrogen prior to DNA/RNA extraction. For genome assembly, genomic DNA was extracted from leaf tissues, and multiple libraries were prepared, including long- and short-read whole genome sequencing (WGS) and Hi-C libraries. For gene annotation, RNA was isolated from both leaf and flower tissues of the same specimen, and both short and long-read RNA-seq libraries were constructed. Long-read WGS was performed using an ONT PromethION sequencer. Short-read WGS, Hi-C, and RNA sequencing (RNA-seq, using cDNA) were conducted using MGI DNBSEQ-T7 platform sequencer with a 150-bp paired-end (insert size 300 bp) sequencing strategy. Additionally, long-read RNA-seq was performed on the same cDNA derived from leaf and flower tissues using the ONT PromethION platform.

Library construction and sequencing were performed by GrandOmics Biosciences (GB, Wuhan, China). DNA/RNA preparation and library generation methods followed protocols described in previous publications[20–22]. The sequencing effort produced approximately 139 Gb of ONT long WGS reads, while MGI platform contributed 114 Gb of short WGS reads and 28 Gb of Hi-C reads (Table 1). RNA-seq reads totaling 69.6 Gb were obtained from both the ONT and MGI sequencing platforms (Table 1).

Data	Number of bases (Gb)	Number of reads
Long WGS read	138.9	167,144,906,028
Short WGS read	113.7	212,936,833,200
Hi-C read	28.1	51,957,833,400
RNA-seq read	69.6	97,852,184,110

Table 1. Sequencing data for *Tigridiopalma magnifica* assembly.

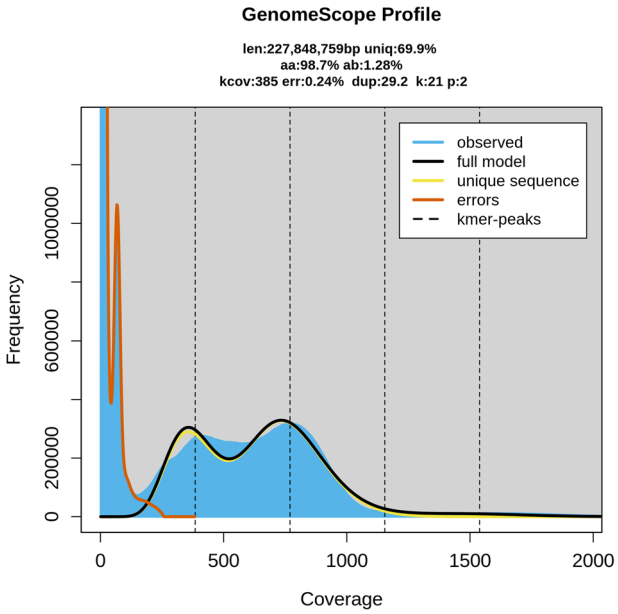

Fig. 2 Genome survey of *Tigridiopalma magnifica* based on 21 *K*-mer analysis.

Data pre-processing. The adapters in the long WGS reads were removed using Porchop v0.2.4[23] with the parameters "--check_reads 100000 --adapter_threshold 20". For the short WGS and Hi-C reads, quality trimming was conducted using fastp v0.23.3[24]. To evaluate genome characteristics, *K*-mer frequency analysis was performed on the short WGS reads. *K*-mer counting was executed using Jellyfish v2.3.0[25] with the parameters -C -m 21 -s 100000000, followed by histogram generation (jellyfish histo -h 1000000). The resulting *K*-mer distribution was analyzed with GenomeScope v2.0[26], using a *K*-mer length of 21 and assuming diploidy (Ploidy = 2) (Fig. 2). The genome size of *T. magnifica* was estimated at 227.85 Mb, with the heterozygosity and duplication were determined to be 1.28% and 29.2% respectively.

Genome assembly. The genome of *T. magnifica* was assembled using Nextdenovo v2.5.2[14], utilizing pre-processed long reads exceeding 10000 bp in length. To eliminate haplotypic duplications within the assembly, Pseudohaploid[27] and Purge_Dups v1.2.6[28] were applied. Scaffhic v1.1[29] was then used to identify misassembly and segment the contigs with trimmed Hi-C reads. Furthermore, to improve the accuracy of the obtained assembly, Inspector v1.3.1[30] was employed to correct the assembly with the long WGS reads. Subsequently, the assembly underwent polishing by Racon v1.5.0[31] with long WGS reads twice, followed by Hapo-G 1.3.8[32] with short WGS reads twice. The telomeres at the contig ends were identified using a telomere identification toolkit (tidk) v0.2.65[33], facilitating error correction in the following scaffolding step. Given that the error rates for ONT reads can reach 20%[34–36], a haplotype-resolved de novo assembly was not attempted. Future efforts should integrate ultra-long and HiFi reads with haplotype phasing workflows, thereby improving superior assembly contiguity and consensus quality.

Hi-C data was processed by HapHiC v1.0.7[15] to perform haplotype-aware genome scaffolding and generate raw contact maps. The scaffolding results were visualized using Juicebox v2.04.06[37], with manual corrections implemented as necessary. Primary corrections included adjustments to chromosome boundaries and the identification of misjoins and inversion errors. Following scaffolding, Rffiller[38] was utilized to enhance genome assemblies by precisely filling gaps in scaffolded assembly. Additional polishing of the gap-closed assembly was performed with Racon and Hapo-G. The telomeres at both ends of each pseudochromosome in the final genome assembly were identified using tidk again (Fig. 3 and Supplementary Table S1). In addition, combining with Hi-C data, we characterized the centromeric regions of the 22 pseudochromosomes using CentIER v3.0[39] to further define the chromosomal architecture. To confirm the consistence, we tried CentIER twice with different parameter in "--matrix1 --matrix2 --bed1 --bed2" which used Hi-C mapping results (Hi-C matrix) obtained from HiC-Pro v3.1.0[40]. These "matrix" and "bed" combination were one with 20000 and 100000, and

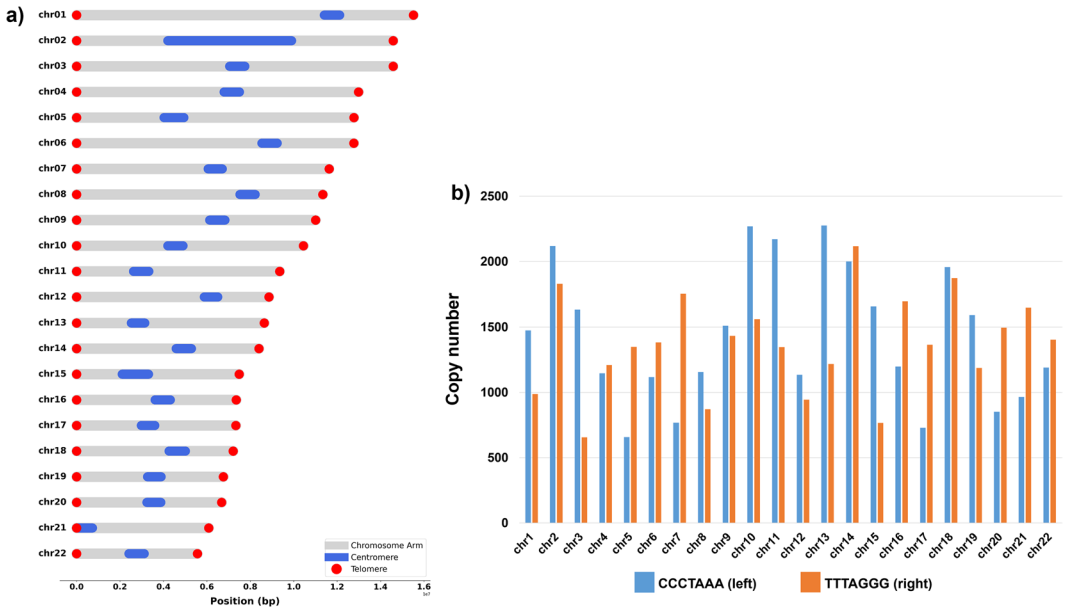

Fig. 3 (**a**) The visualization of telomeres (at both termini) and centromeres in the 22 pseudochromosomes. (**b**) Copy number of telomeres in all pseudochromosome of *T. magnifica*.

the other with 40000 and 20000.The initial assembly generated by Nextdenovo measured 257,764,685 bp, with an N50 of 8,641,021 bp (Table 2). After performing Pseudohaploid, the assembly length was 256,818,587 bp. This was further streamlined to 217,555,391 bp using Purge_Dups to eliminate allelic duplicates, and reached 217,457,672 bp after refinement with Inspector. After scaffolding with Hi-C reads, the assembly size decreased to 217,339,795 bp, achieving an N50 of 11,008,659 bp, and was organized into 22 pseudochromosomes (Fig. 4). The largest chromosome measured 15,514,458 bp while the shortest was 5,564,566 bp (Supplementary Table S2). Utilizing bedtools v2.31.1[41], the average GC content determined for the entire genome was 40.73%, while the content for discrete chromosomes varied from 36.03% to 55.10%.

While the assembly achieves chromosomal-level continuity, two remaining intra-chromosomal gaps, one on chromosome 5 and one on chromosome 15, were distinguished. Telomeres at both ends of all the pseudochromosomes were identified (Fig. 3a) and the statistics of telomeric repeats were shown in Fig. 3b and Supplementary Table S1. Two running of CentIER gave the same results, and centromeres regions were identified in all pseudochromosomes which spaned from 620,000 bp to bp 5,699,999 bp. Notably, the centromere on chromosome 2 exhibited a significantly large span, while the centromeric region of chromosome 21 demonstrated close to an end side of the chromosomes. Future studies should utilize HiFi and ultra-long reads to refine these boundaries.

Repeat and gene annotation. The repetitive regions of the *T. magnifica* genome were analyzed using EDTA v2.0.1[42] and RED v2.0[43]. The analysis revealed that repetitive sequences constituted 21.60% (EDTA) and 29.57% (RED) of the genome. Long terminal repeats (LTRs) represented the most abundant repetitive sequences, accounting for 14.07% (30,573,313 bp) of the genome, while terminal inverted repeats (TIRs) comprised 3.23% (6,988,191 bp) of the genome (Table 3). Unknown LTRs dominated the repetitive landscape, constituting 7.85% (17,060,575 bp) of the total genome size. The combined analysis using EDTA and RED indicated that 77,135,077 bp (35.49%) of the genome was soft-masked.

Gene prediction was conducted using BRAKER3 v3.0.8[44] and the Funannotate pipeline v1.8.16[45]. The reference proteins utilized for gene prediction are specified in Table 4. After prediction, the functional annotation process of genes incorporated multiple databases, including dbCAN v13.0[46], EggNOG v5.0.2[47], Gene Ontology (GO)[48], Kyoto Encyclopedia of Genes and Genomes (KEGG)[49], InterPro v5.75-106.0[50], MEROPS v12.0[51], Pfam v37.4[52], SignalP 5.0b[53] and UniProt v2025_03[54]. The analysis identified 43,291 protein-coding genes, which encode 48,218 proteins. Through functional annotation, 43,034 genes (99.41%) were successfully annotated in at least one database (Table 5). Using predicted genes, synteny blocks in the genome were identified using jcvi v1.3.841[55]. The syntenic relationship, combining with GC, repeat and gene contents were all illustrated in the circos plot (Fig. 4a) made by Shinycircos v2.0[56]. These comprehensive genomic features have elucidated previously undocumented structural and functional characterization of *T. magnifica* genome (Fig. 4, Supplementary Table S1, Table S2 and Supplementary Fig. S1–S3).

Finally, a K_s analysis was conducted using wgd v1.1.2[57] to investigate potential whole-genome duplication (WGD) events in *T. magnifica*. The result indicated no recent WGD events or signatures of ancient polyploidy happened (Supplementary Figure S4).

Contig statistics of initial assembly using Nextdenovo	
The length of sequence (bp)	**The order of sequence length**
N10 = 14571790	L10 = 2
N20 = 12761392	L20 = 4
N30 = 11016523	L30 = 6
N40 = 10127962	L40 = 9
N50 = 8641021	L50 = 12
N60 = 7334182	L60 = 15
N70 = 7202994	L70 = 18
N80 = 6098122	L80 = 22
N90 = 3816323	L90 = 28
N100 = 29912	L100 = 37
Total length	257,764,685 bp
Average length	3,682,352 bp
Largest length	15,514,762 bp
Minimum length	29,912 bp
Scaffold statistics after Hi-C scaffolding	
N10 = 14579561	L10 = 2
N20 = 14578959	L20 = 3
N30 = 12774436	L30 = 5
N40 = 11631895	L40 = 7
N50 = 11008659	L50 = 9
N60 = 9354070	L60 = 11
N70 = 8639881	L70 = 13
N80 = 7350993	L80 = 16
N90 = 6756761	L90 = 19
N100 = 5564566	L100 = 22
Total length	217,339,795 bp
Average chromosome length	9,879,081.6 bp
Largest chromosome length	15,514,458 bp
Minimum chromosome length	5,564,566 bp

Table 2. Statistic of *Tigridiopalma magnifica* genome assembly.

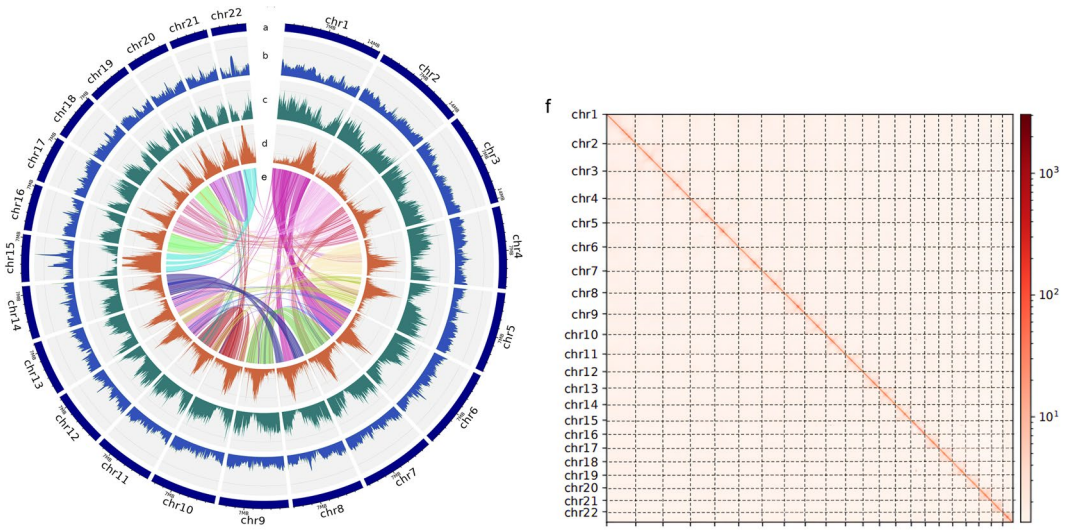

Fig. 4 Characteristics of the *Tigridiopalma magnifica* genome. (**a**–**e**) Circular tracks represent, from outer to inner, 15 chromosomal-level scaffolds (Chr1 - Chr22), GC content (36.03%–55.10%), gene density (0–710), percentage of repeats (8.01%–100%), and sequence synteny within the genome, respectively (window size = 90 kb). (**f**) The Hi-C contact map of the *Tigridiopalma magnifica* genome. chr 01–22 represented for the 22 pseudochromosomes.

Repeat class	Repeat family	Number of elements	Total length (bp)	Percentage in the assembly
LTR		41,366	30,573,313	14.07%
	Copia	3,543	2,842,801	1.31%
	Gypsy	13,149	10,669,937	4.91%
	unknown	24,674	17,060,575	7.85%
TIR		14,298	6,988,191	3.23%
	CACTA	2,645	949,970	0.44%
	Mutator	9,088	4,755,869	2.19%
	PIF_Harbinger	789	402,998	0.19%
	Tc1_Mariner	439	126,395	0.06%
	hAT	1,337	752,959	0.35%
nonLTR		149	22,575	0.01%
	pararetrovirus	149	22,575	0.01%
nonTIR		107,273	3,731,245	1.72%
	helitron	8,106	3,731,245	1.72%
repeat fragment		18,234	4,781,811	2.20%
Total		83,683	46,947,974	21.60%

Table 3. Repeat content of the Hi-C scaffolded assembly by EDTA.

Species	GenBank accession number or website
Corymbia_citriodora_subsp. variegata	GCA_014858505.1
Eucalyptus grandis	GCF_016545825.1
Melastoma candidum	GCA_023653495.1
Melastoma dodecandrum	https://doi.org/10.1016/j.jgg.2021.10.004
Punica granatum	GCF_007655135.1
Rhodamnia argentea	GCF_020921035.1
Syzygium oleosum	GCF_021117445.2
Trapa natans	GCA_035582545.1

Table 4. Protein sequences of the species used for gene prediction.

Annotation database	Number of genes	Percentage (%)
dbCAN	1,531	3.54
EggNOG	39,472	91.07
KEGG	4,163	9.62
GO	31,164	71.99
InterPro	38,791	89.61
MEROPS	1,429	3.30
Pfam	30,690	70.89
SignalP	3,670	8.48
UniProt	11,625	26.85
Total	43,034	99.41

Table 5. Summary of gene functional annotations in the *Tigridiopalma magnifica* genome across multiple databases.

Data Records

The sequenced reads are accessible in the National Center for Biotechnology Information (NCBI) Sequence Read Archive, with the following accession numbers: SRR35279824[58] for the long WGS reads, SRR35279823[59] for the short WGS reads, SRR35279821[60] for the Hi-C reads, and SRR35279822[61] for the RNA-seq reads. *Tigridiopalma magnifica* final genome assembly, annotation, and protein coding sequences were submitted to Figshare[62]. The assembled genome was also deposited in the NCBI under the accession number JBJJNS000000000[63].

Technical Validation

To evaluate the assembly quality, BUSCO v5.5.0[16] was implemented using the eudicots_odb10.2020-09-10 database, which contains 2,326 conserved core genes specific to eudicots. The analysis demonstrated that the genome completeness of *T. magnifica* reached 95.0%, comprising 70.2% complete single-copy genes, 24.8% complete duplicated genes, 0.8% fragmented genes, and 4.2% missing genes.

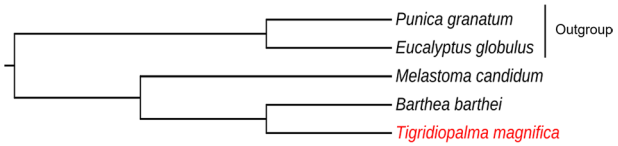

Fig. 5 Phylogenomic placement of *Tigridiopalma magnifica* based on single-copy orthologous proteins. Branch lengths represent substitutions per site.

Assembly quality assessment also utilized AssemblyQC v2.1.1[64]. The AssemblyQC generated the Long Terminal Repeat Assembly Index (LAI), which evaluates assembly contiguity based on repetitive sequences[17], along with *K*-mer-based metrics for accuracy and completeness through Merqury v1.3[18]. Various sequencing read types, including long-read WGS, short-read WGS, and RNA-seq, were aligned to the reference assembly using Minimap2 v2.29[65], BWA v0.7.19[66], and HISAT2 v2.2.1[67], respectively. Alignment success rates (mapping ratios) were calculated using Samtools v1.21[68].

LAI analysis yielded a high score of 16.01. Merqury indicated the quality value (QV) score was 42.09. The mapping-based evaluation of read alignments to the reference assembly demonstrated the following rates: 97.41% and 94.75% for long- and short-read WGS data, respectively; 97.56% and 96.32% for short RNA-seq reads from leaf and flower tissues, respectively; 84.95% and 87.62% for long RNA-seq reads from leaf and flower tissues, respectively.

To validate the taxonomic identity of the *T. magnifica* assembly, we performed a phylogenomic reconstruction using OrthoFinder v2.5.5[69,70] alongside two Melastomataceae references and two outgroup species, *Punica granatum* and *Eucalyptus globulus*. The resulting species tree (Fig. 5) demonstrates that *T. magnifica* is most closely related to *B. barthei*, consistent with their shared placement in the Sonerileae tribe[71].

Data availability

The raw sequencing reads supporting the findings of this study are publicly available in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1321007 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1321007). The long RNA sequencing data (SRA run accession: SRR35279822) comprise four files in FASTQ format: two derived from leaf tissue (20231109-NPL2300541-P6-PAS12376-fast_barcode86_1.pass.fastq.gz and 20231109-NPL2300541-P6-PAS12376-fast_barcode86_2.pass.fastq.gz) and two derived from flower tissue (20231109-NPL2300541-P6-PAS12376-fast_barcode87_1.pass.fastq.gz and 20231109-NPL2300541-P6-PAS12376-fast_barcode87_2.pass.fastq.gz). The corresponding short RNA-seq reads consist of two paired-end FASTQ files from leaf tissue (E250003330_L01_292_1.fq.gz and E250003330_L01_292_2.fq.gz) and two from flower tissue (E200014034_L01_377_1.fq.gz and E200014034_L01_377_2.fq.gz). The website https://doi.org/10.6084/m9.figshare.30336466 contains four files. One is *T. magnifica* assembled genome under the file name of "T.magnifica genome assembly.fasta". The remaining three files are gene prediction and annotation under the file names of "T.magnifica.gff3", "T.magnifica.cds-transcripts.fasta", and "T.magnifica.proteins.fasta".

Code availability

All software and pipelines utilized in this study were executed following the guidelines of the published tools. The parameters and version numbers of software and databases are specified in the Methods section. Any elements not detailed in the Methods were implemented using default parameters. No custom scripts were utilized.

Received: 2 November 2025; Accepted: 25 March 2026;
Published online: 01 April 2026

References

1. Xu, W.-B. *et al.* Human activities have opposing effects on distributions of narrow-ranged and widespread plant species in China. *Proc. Natl. Acad. Sci. USA* **116**, 26674–26681 (2019).
2. Theissinger, K. *et al.* How genomics can help biodiversity conservation. *Trends Genet.* **39**, 545–559 (2023).
3. Garner, B. A. *et al.* Genomics in conservation: case studies and bridging the gap between data and application. *Trends Ecol. Evol.* **31**, 81–83 (2016).
4. Chen, J. A new genus *Tigridiopalma* C. Chen of Malastomataceae. *Acta Bot. Yunnan.* **1**, 106–108 (1979).
5. Ren, H. *et al.* Reintroduction of *Tigridiopalma magnifica*, a rare and critically endangered herb endemic to China. *Oryx* **46**, 391–398 (2012).
6. *List of key protected plants in China*, https://www.gov.cn/zhengce/zhengceku/2021-09/09/content_5636409.htm (Accessed 17 September 2025).
7. Ulloa Ulloa, C. *et al.* Melastomataceae: global diversity, distribution, and endemism. in *Systematics, Evolution, and Ecology of Melastomataceae* (eds. Goldenberg, R., Michelangeli, F. A. & Almeda, F.) 3–28 (Springer International Publishing, 2022).
8. Zhong, Y. *et al.* Chromosomal-level genome assembly of *Melastoma candidum* provides insights into trichome evolution. *Front. Plant Sci.* **14**, 1126319 (2023).
9. Huang, W. *et al.* De novo genome assembly and population genomics of a shrub tree *Barthea barthei* (Hance) krass provide insights into the adaptive color variations. *Front. Plant Sci.* **15**, 1365686 (2024).
10. Reginato, M., Neubig, K. M., Majure, L. C. & Michelangeli, F. A. The first complete plastid genomes of Melastomataceae are highly structurally conserved. *PeerJ* **4**, e2715 (2016).
11. Zheng, X., Ren, C., Huang, S., Li, J. & Zhao, Y. Structure and features of the complete chloroplast genome of *Melastoma dodecandrum*. *Physiol. Mol. Biol. Plants* **25**, 1043–1054 (2019).
12. Zhou, Q., Ng, W. L., Wu, W., Zhou, R. & Liu, Y. Characterization of the complete chloroplast genome sequence of *Tigridiopalma magnifica* (Melastomataceae). *Conserv. Genet. Resour.* **10**, 571–573 (2018).

13. Garg, V. *et al*. Unlocking plant genetics with telomere-to-telomere genome assemblies. *Nat. Genet.* **56**, 1788–1799 (2024).
14. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107, https://doi.org/10.1186/s13059-024-03252-4 (2024).
15. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
16. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
17. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126, https://doi.org/10.1093/nar/gky730 (2018).
18. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
19. Zeng, S. J., Duan, J. & Li, L. N. Plant regeneration from leaf explants of *Tigridiopalma magnifica* (Melastomataceae). *Pak. J. Bot.* **40**, 1179–1184 (2008).
20. Wang, Z.-F. *et al*. Genome assembly of *Musa beccarii* shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes. *GigaScience* **12**, giad005 (2023).
21. Wang, Z.-F. *et al*. Chromosome-level genome assembly and demographic history of Euryodendron excelsum in monotypic genus endemic to China. *DNA Res.* **31**, dsad028 (2024).
22. Wang, Z.-F. *et al*. Chromosome-scale assemblies of three *Ormosia* species: repetitive sequences distribution and structural rearrangement. *GigaScience* **14**, giaf047, https://doi.org/10.1093/gigascience/giaf047 (2025).
23. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132, https://doi.org/10.1099/mgen.0.000132 (2017).
24. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
25. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
26. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
27. Schatz, M. C. *Pseudohaploid. GitHub* https://github.com/schatzlab/pseudohaploid (2020).
28. Guan, D. F. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
29. *Scaffhic v1.1. GitHub* https://github.com/wtsi-hpag/scaffHiC (2022).
30. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
31. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746, https://doi.org/10.1101/gr.214270.116 (2017).
32. Aury, J. M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom. Bioinform.* **3**, lqab034, https://doi.org/10.1093/nargab/lqab034 (2021).
33. Brown, M. R., De La Rosa, P. G. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaf049 (2025).
34. Chen, Y. *et al*. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
35. Mochizuki, T. *et al*. A practical assembly guideline for genomes with various levels of heterozygosity. *Brief. Bioinform.* **24**, bbad370 (2023).
36. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).
37. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258.e1, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
38. Midekso, F. D. & Yi, G. RFfiller: a robust and fast statistical algorithm for gap filling in draft genomes. *PeerJ* **10**, e14186 (2022).
39. Xu, D. *et al*. CentIER: accurate centromere identification for plant genome. *Plant Commun.* **5**, 101046 (2024).
40. Servant, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C processing. *Genome Biol.* **16**, 259 (2015).
41. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
43. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 227, https://doi.org/10.1186/s12859-015-0654-5 (2015).
44. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777, https://doi.org/10.1101/gr.278090.123 (2024).
45. Palmer, J. M. & Stajich, J. Funannotate v1.8.1: Eukaryotic genome annotation. *Zenodo* https://doi.org/10.5281/zenodo.1134477 (2020).
46. Zhang, H. *et al*. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101, https://doi.org/10.1093/nar/gky418 (2018).
47. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
48. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338, https://doi.org/10.1093/nar/gky1055 (2019).
49. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
50. Blum, M. *et al*. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res.* **53**, D444–D456, https://doi.org/10.1093/nar/gkae1082 (2025).
51. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**, D227–D233, https://doi.org/10.1093/nar/gkp971 (2010).
52. Paysan-Lafosse, T. *et al*. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.* **53**, D523–D534, https://doi.org/10.1093/nar/gkae997 (2025).
53. Almagro Armenteros, J. J. *et al*. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423, https://doi.org/10.1038/s41587-019-0036-z (2019).
54. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617, https://doi.org/10.1093/nar/gkae1010 (2025).
55. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
56. Wang, Y. *et al*. shinyCircos-V2.0: Leveraging the creation of Circos plot with enhanced usability and advanced features. *iMeta* **2**, e109 (2023).

57. Chen, H. & Zwaenepoel, A. Inference of ancient polyploidy from genomic data. in *Polyploidy: Methods and Protocols* (eds. Van de Peer, Y.) 3–18 (Springer US, 2023).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35279824 (2025).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35279823 (2025).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35279821 (2025).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35279822 (2025).
62. *Figshare* https://doi.org/10.6084/m9.figshare.30336466 (2025).
63. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054741195.1 (2026).
64. Rashid, U. *et al*. AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. *Bioinformatics* **40**, btae477, https://doi.org/10.1093/bioinformatics/btae477 (2024).
65. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
67. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
68. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
69. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
70. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
71. Zhou, Q. *et al*. Analyses of plastome sequences improve phylogenetic resolution and provide new insight into the evolutionary history of Asian Sonerileae/Dissochaeteae. *Front. Plant Sci.* **10**, 1477 (2019).

Acknowledgements

This work was supported by Science & Technology Fundamental Resources Investigation Program (Grant No. 2022FY100500).

Author contributions

Z.F.W. and X.J.G. supervised the study. Z.F.W. designed the method and validated the data. Z.F.W., X.J.G., and T.W.X. revised the manuscript. D.Q.V. performed the data analyses and wrote the manuscript. All authors approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07127-0.

Correspondence and requests for materials should be addressed to Z.-F.W. or X.-J.G.