scientific data

OPEN

DATA DESCRIPTOR

Chromosome level genome assembly of taro (*Colocasia esculenta*)

Yalin Sun[1,4], Zhixin Wang[1,4], Haining Wang[1], Ying Wang[1], Ziyan Qian[1], Yuping Liu[1], Yingnan Yang[1], Xuxiong Qin[2], Gang Huang[3], Xiaobo Ying[3] ✉, Gang Song[2] ✉ & Honglian Zhu[1] ✉

Taro (*Colocasia esculenta*), a crop native to Southeast Asia, is classified under the genus *Colocasia* of the family Araceae. Here, the taro genome was assembled to the chromosome level, yielding a 2,336.63 Mb assembly with a scaffold N50 of 181.70 Mb. A total of 96.86% of assembled contigs were positioned onto 14 chromosomes through Hi-C scaffolding. Repeat annotation indicated that 85.16% of the genome (1,989.98 Mb) comprises repetitive sequences. A total of 28,253 protein-coding genes were predicted, with 95.14% assigned functional annotations. In addition, 2,842 rRNAs, 1,392 tRNAs, 452 snRNAs, and 68 miRNAs were identified. The availability of this reference genome lays the groundwork for comprehensive genomic analyses and facilitates genetic improvement in taro.

Background & Summary

Taro (*Colocasia esculenta*), belonging to the family Araceae, is the world's fifth-ranked tuberous root crop cultivated extensively across Asia, Africa, Oceania, and the Americas[1]. Taro is probably one of the oldest crops and exhibits remarkable phenotypic and genetic diversity, particularly in corm morphology, pigmentation, and tolerance to abiotic stresses[2]. The diversity of taro cultivars enables successful cultivation across a wide range of environments, from wetlands to rain-fed upland in tropical and subtropical regions[3]. In the face of growing climate instability and the deepening crisis of global food security, taro is increasingly being recognized as a high-yielding and climate-resilient crop for sustainable agriculture[4].

The increasing economic importance of taro has heightened the demand for high-quality genome. A chromosome-level genome of taro was reported. The assembled genome size was approximately 2.4 Gb, but comprised more than 7000 scaffolds[5]. Complete mitochondrial and chloroplast genomes have been sequenced, providing insights into organelle structure, gene content, and evolutionary dynamics[6,7]. However, a high-quality genome is still required to fully elucidate taro's evolutionary history, population genomics, and to facilitate marker-assisted breeding for improved yield, disease resistance, and environmental adaptability.

Among the diverse landraces of taro, 'Bun Long' taro stands out as one of the most important cultivars in China[8]. It is widely appreciated for its dense texture and rich aroma when cooked. Owing to the unique quality traits[9], 'Bun Long' represents an important genetic resource for taro breeding. Here, to provide a more comprehensive genomic resource for 'Bun Long' taro, we established a chromosome-level genome assembled through PacBio SMRT[10] and Hi-C sequencing[11]. The assembly displays excellent completeness, sequence continuity, and structural correctness. Genome annotation revealed the landscape of transposable elements, predicted protein-coding genes, and associated Gene Ontology (GO) and KEGG functional pathways. This reference genome represents a high-contiguity assembly of *Colocasia esculenta* and serves as a foundational resource for comparative genomics, evolutionary studies, and trait discovery within the Araceae family. The genome will accelerate identification of key genes for yield, stress tolerance, and corm quality, supporting taro breeding and improvement efforts.

[1]Institue of vegetables, Wuhan academy of agricultural sciences, Wuhan, China. [2]Nanning Guifuyuan Agriculture Co., Ltd. Nanning, Guangxi Zhuang Autonomous Region, Guangxi, China. [3]HuaZhi Biotechnology Co., Ltd, Changsha, China. [4]These authors contributed equally: Yalin Sun, Zhixin Wang. ✉e-mail: hawk72_ying@163.com; 3342192947@qq.com; zhuhonglian2014@163.com

Fig. 1 Photographs illustrating the morphology of *Colocasia esculenta*.

Methods

Sample collection, DNA extraction and sequencing. For this study, plant material of 'Bun Long' taro was sourced from Qingshan Town, Lipu, Guangxi Zhuang autonomous region, China. DNA for sequencing was obtained from the leaves of *C. esculenta* 'Bun Long' taro (Fig. 1) using CTAB method. DNA was quantified with the Quant-iT PicoGreen assay (Invitrogen) and evaluated for purity using 260/280 nm absorbance on a UV spectrophotometer.

Following quality and concentration assessment, genomic DNA was isolated with the PacBio Nanobind® Plant Nuclei Kit TR (Pacific Biosciences, Menlo Park, CA, USA) in accordance with the supplier's guidelines. Single-molecule real-time (SMRT) libraries (15-kb) were prepared with the PacBio protocol and sequenced on the PacBio Revio platform, generating ~80.77 Gb (34.6×) of filtered data. Hi-C libraries were generated with the restriction enzyme MboI; biotinylated DNA was captured with streptavidin-coated magnetic beads, followed by end repair, A-tailing, adapter ligation, PCR cycle optimization, and purification. Libraries passing quality control were pooled by effective concentration and target output, and sequenced on the BGI PE150 platform, yielding 224.12 Gb (95.9×) of raw data (Table 1). All sequencing was performed by HuaZhi Biotechnology Co., Ltd.

Genome survey and assembly. K-mer analyses of the HiFi reads, performed with Jellyfish v2.3.0[12] and GenomeScope2[13], estimated a genome size of 2.37 Gb, heterozygosity of 1.83%, and repeat content of 74.8% (Fig. 2).

For the *de novo* genome assembly of *C. esculenta*, raw data were first converted to FASTQ format using the bam2fastq v3.0.0 tool pipeline (https://github.com/PacificBiosciences/pbtk#bam2fastx). A primary assembly was then generated with hifiasm v0.19.5[14] using default parameters. The resulting draft genome consisted of 499 contigs, totaling 2.337 Gb, with a contig N50 of 18.02 Mb and a GC content of 42.55% (Table 2).

Filtered Hi-C reads, obtained after adapter and quality trimming, were aligned to the primary assembly using Juicer v1.6[15] with default settings to produce a deduplicated read set. Then, 3D-DNA v180922[16] was used to anchor primary contigs to pseudochromosomes. The resulting assembly was manually curated using Juicerbox[17] by visually inspecting Hi-C interaction heatmaps to correct misjoins and reorder/orient contigs. This process yielded fourteen pseudochromosomes, ranging from 105.4 Mb to 211.4 Mb, based on distinct interaction signals (Fig. 3). Consequently, the *C. esculenta* genome was assembled to the chromosome level, with a total size of 2.263 Gb (Table 3).

Annotation of repetitive sequences. Repeat sequences were identified using two approaches: homolog-based and *ab initio*. In the homolog-based approach, known repeats were predicted by aligning the genome to the RepBase database with RepeatMasker v4.1.2[18] and RepeatProteinMask Revision 1.36[18]. For the *ab initio* approach, a *de novo* repeat library was generated using RepeatModeler v2.0.1[19] and LTR-FINDER v1.0.7[20], followed by RepeatMasker prediction. Tandem repeats were additionally detected with TRF v4.09.1[21].

Libraries	Sequencing Platform	Total Base(Gb)	Sequence coverage (×)	Reads Numbers	Reads GC(%)
PacBio	PacBio Revio	80.77	34.6	4,590,360	41.8
Hi-C	BGI PE150	224.12	95.9	739,817,274	41.78

Table 1. Information of sequencing data generated for the genome assembly of *Colocasia esculenta*.

Assembly characteristics	Values
Number of contigs	499
Assembly length(bp)	2,336,626,558
GC(%)	43
Contig N50(bp)	18,022,072
anchor ratio(%)	96.8585
HiFi reads mapping rate(%)	99.96
HiFi reads coverage(%)	99.99
BUSCO(%)	96.9
KAT(%)	99.68
LTR assembly index	16.32
QV	63.2421

Table 2. Statistics of the *Colocasia esculenta* genome assembly and scaffolding.

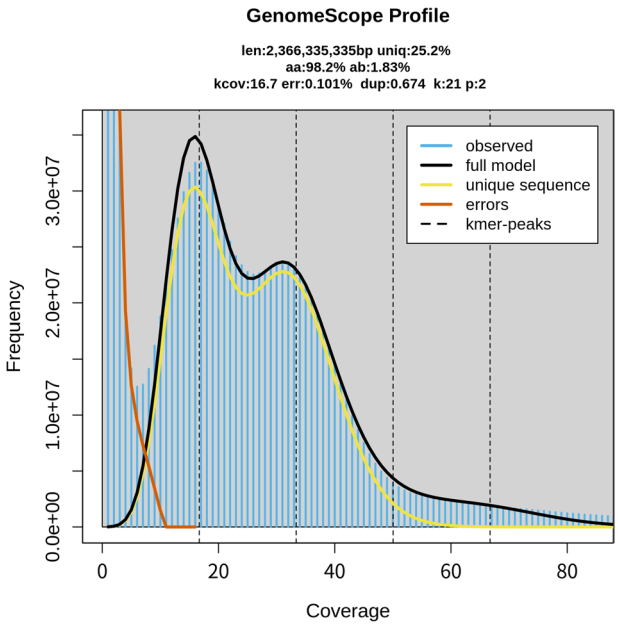

Fig. 2 K-mer analysis of the *Colocasia esculenta* genome based on the 19-mer frequency distribution.

The genome contained 1.99 Gb of repeat elements, representing 85.16% of the *C. esculenta* genome (Table 4). Long terminal repeats (LTRs) were the predominant type, representing 62.44% of these elements (Table 5). The genome assembly quality was further supported by a LTR Assembly Index (LAI) of 16.32, calculated with LTR_retriever v2.9.0[22] (Table 2).

Gene prediction and annotation. Gene annotation was performed on the repeat-masked genome. A comprehensive approach was employed to predict protein-coding genes, integrating transcriptome-based annotations, homology-based inference, and *de novo* computational predictions. Genome-aligned RNA-seq assemblies from the CNCB database (CRA028667)[23] were used as transcriptome evidence. Reads from RNA-seq were mapped to the reference genome with HISAT2 v2.1.0[24]. Transcripts were assembled with Cufflinks v2.2.1[25] using default parameters. All sample transcripts were merged and subjected to protein-coding prediction and quality control via TransDecoder in PASA v2.4.1[26]. Only intact transcripts were used for further analyses. The protein sequences from three members of Araceae family—*Amorphophallus konjac*[27], *Zantedeschia elliottiana*[28], *Pistia stratiotes*[29] were aligned to the genome of *C. esculenta* using Exonerate v2.4.0[30] to identify high-quality protein structures. Augustus v3.4.0[31] and GlimmerHMM v3.0.4[32] were used for *de novo* gene prediction, both trained on the high-quality transcripts obtained in the previous step, following the software manuals. Predicted gene sets

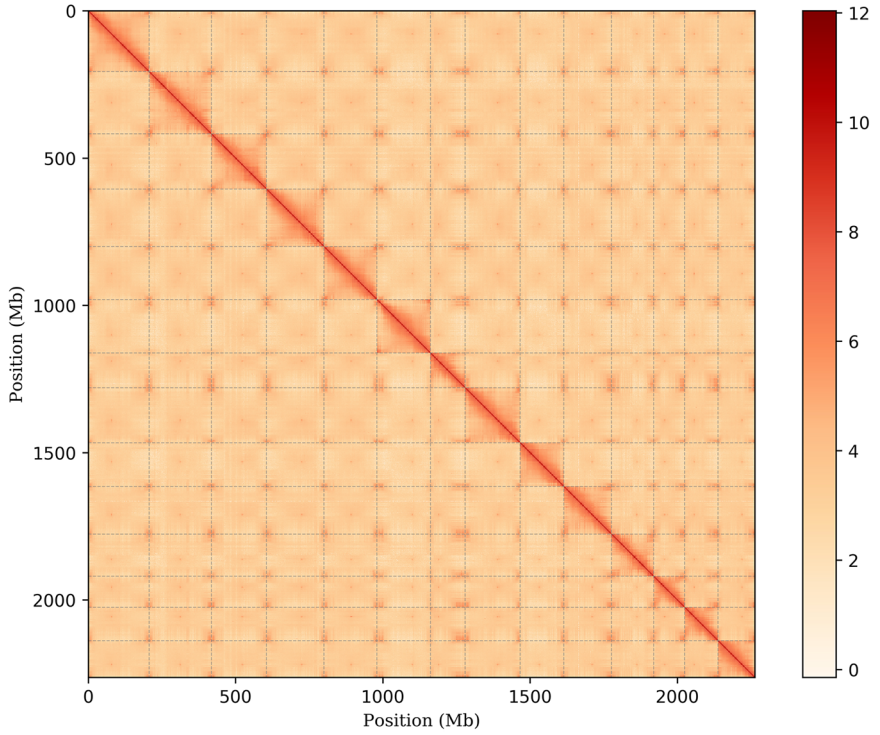

Fig. 3 Hi-C contact map at chromosome-scale for the genome assembly of *Colocasia esculenta*.

Chromosome	Number of Contigs	Length of Superscaffold (bp)
Chr1	19	206,526,311
Chr2	33	211,375,948
Chr3	26	187,053,525
Chr4	18	195,361,014
Chr5	21	179,761,935
Chr6	62	181,704,944
Chr7	24	117,894,572
Chr8	19	186,956,590
Chr9	33	148,221,615
Chr10	12	161,672,056
Chr11	40	143,252,809
Chr12	15	105,394,347
Chr13	25	113,431,070
Chr14	24	124,793,446
TOTAL	371	2,263,400,182

Table 3. Statistical of Colocasia esculenta genomic in chromosome distribution.

from the above approaches were consolidated into a non-redundant, more complete set using MAKER v3.01.03[33], followed by gene structure refinement with PASA incorporating transcriptome data. We identified and annotated 28,253 protein-coding genes, averaging 11,527 bp in length (Table 6). Gene density, GC content, and the distribution of repetitive sequences across each chromosome were visualized using TBtools[34] (Fig. 4).

Protein-coding genes were functionally annotated by aligning their sequences against the InterPro[35], NR[36], GO[37], KEGG[38], EggNOG[39], SwissProt[40], and TrEMBL[40] databases using DIAMOND v2.0.7[41] (-c 1 -f 5 -e 1e-5–very-sensitive -k 10). Among the 26,881 protein-coding genes identified, a majority (95.14%) were annotated with at least one functional term (Table 7).

For annotation of noncoding RNA, tRNA genes were identified based on their structural features using tRNAscan-SE v2.0.9[42]. rRNA genes were predicted with RNAmmer v1.2[43]. Additionally, miRNA and snRNA sequences were predicted using covariance models from the Rfam database[44] via the INFERNAL v1.1.4 software[45]. Finally, a total of 4,754 noncoding RNAs were predicted (Table 8), including 2,842 ribosomal RNAs (rRNAs), 1,392 transfer RNAs (tRNAs), 452 small nuclear RNAs (snRNAs), and 68 microRNAs (miRNAs).

Type	Repeat Size (bp)	% of genome
TRF	165650807	7.09
Repeatmasker	451070794	19.3
RepeatProteinmask	447332493	19.14
de novo	1951789625	83.52
Total	1989981140	85.16

Table 4. Statistical of repetitive elements predicted in the assembled genome of *Colocasia esculenta* based on the different methods used.

Type	DNA	LINE	SINE	LTR	Other	Unknown	Total TE
RepeatMasker TEs Length (bp)	45,262,243	23,020,103	175,095	386,489,158	1,041	484,891	449,663,431
RepeatMasker TEs % in genome	1.94	0.99	0.01	16.54	0.00	0.02	19.24
RepeatProteinMask TEs Length (bp)	2,143,706	1,356,669	0	443,832,848	0	0	447,332,493
RepeatProteinMask TEs % in genome	0.09	0.06	0	18.99	0	0	19.14
de novo Length (bp)	400,484,594	96,773,218	3,541,589	1,442,699,247	0	69,289,553	1,934,090,669
de novo % in genome	17.14	4.14	0.15	61.74	0.00	2.97	82.77
Combined TEs Length (bp)	416,520,680	105,817,484	3,629,083	1,459,082,182	1,041	69,770,401	1,957,741,551
Combined TEs% in genome	17.82	4.53	0.16	62.44	0	2.99	83.78

Table 5. Information on the repeat annotation in the *Colocasia esculenta* genome assembly.

Method/Software	Number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length(bp)	Average intron length(bp)
de novo/AUGUSTUS	52,417	15,504.09	1,136.75	4.28	265.66	4,381.61
de novo/GlimmHMM	141,724	15,215.85	529.80	3.37	157.32	6,203.07
homo/*A. konjac*	54,964	8,899.88	1,051.53	3.52	298.68	3,113.62
homo/*Z. elliottiana*	31,385	10,185.15	1,022.25	4.00	255.82	3,058.34
Homo/*P. stratiotes*	18,531	12,181.96	1,233.84	5.09	242.60	2,679.46
trans.orf/RNAseq	12,062	11,911.57	1,149.51	5.55	356.23	2,184.42
MAKER	36,366	11,216.98	1,137.58	4.40	309.15	2,899.43
PASA	28,253	11,527.20	1,261.17	4.99	319.24	2,477.38

Table 6. Number of gene annotated in the assembled genome of **Colocasia esculenta** based on the different methods used.

Data Records

All raw whole-genome data have been deposited in the National Genomics Data Center (NGDC) under BioProject PRJCA042833. PacBio, Hi-C, and RNA-seq data are available in the Genome Sequence Archive (CRA028667), and the genome assembly in the Genome Warehouse (GWHGEVN00000000.1). The same sequencing datasets and the assembly are deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1305327 SRP608616[46] and GenBank under the accession JBQGWC000000000[47].

The corresponding BioSample is registered under the accession SAMN50493110. The dataset comprises multiple sequencing libraries, each assigned a distinct SRA accession number: long-read genomic sequencing performed on the PacBio Revio platform (single-end) is deposited under accession SRR34972530; Hi-C data generated on the DNBSEQ platform (paired-end) is available under accession SRR34972529; and full-length transcriptome sequencing conducted on the PacBio Revio platform (single-end) is archived under accession SRR34972528.

The data of the genome annotations, along with associated predicted coding sequences, and protein sequences have been deposited at Figshare[48]. Gene structural annotations are provided in the GFF3 file "Colocasia_esculenta.Genome.V1.gff3". Predicted CDS and protein sequences are respectively available in FASTA format files "Colocasia_esculenta.Genome.V1.cds" and "Colocasia_esculenta.Genome.V1.pep". Additionally, annotations of repetitive elements are compiled in the GFF3 file "all.gff3".

Technical Validation

To assess the completeness of the genome assembly, BUSCO v5.2.2[49] was applied to search for conserved single-copy orthologs present in the embryophyta_odb10 dataset. The results of the BUSCO assessment showed that the assembled genome included 1,564 complete single-copy orthologs, representing 96.9% completeness. (Table 9). An LAI of 16.32 was obtained, indicating that the *C. esculenta* genome assembly meets the standards of a reference-quality genome (Table 2). The assembly quality of the was also assessed using KAT v2.4.2[50] and Merqury v1.3.43[51]. KAT showed high accuracy, with 99.68% of the 1.032×10^9 distinct K-mers in the assembly supported by raw reads. Merqury estimated a consensus quality value (QV) of 63.2 based on HiFi k-mers (Table 2). Using Minimap2 v2.21[52] to assess assembly accuracy, PacBio long reads were remapped to the final

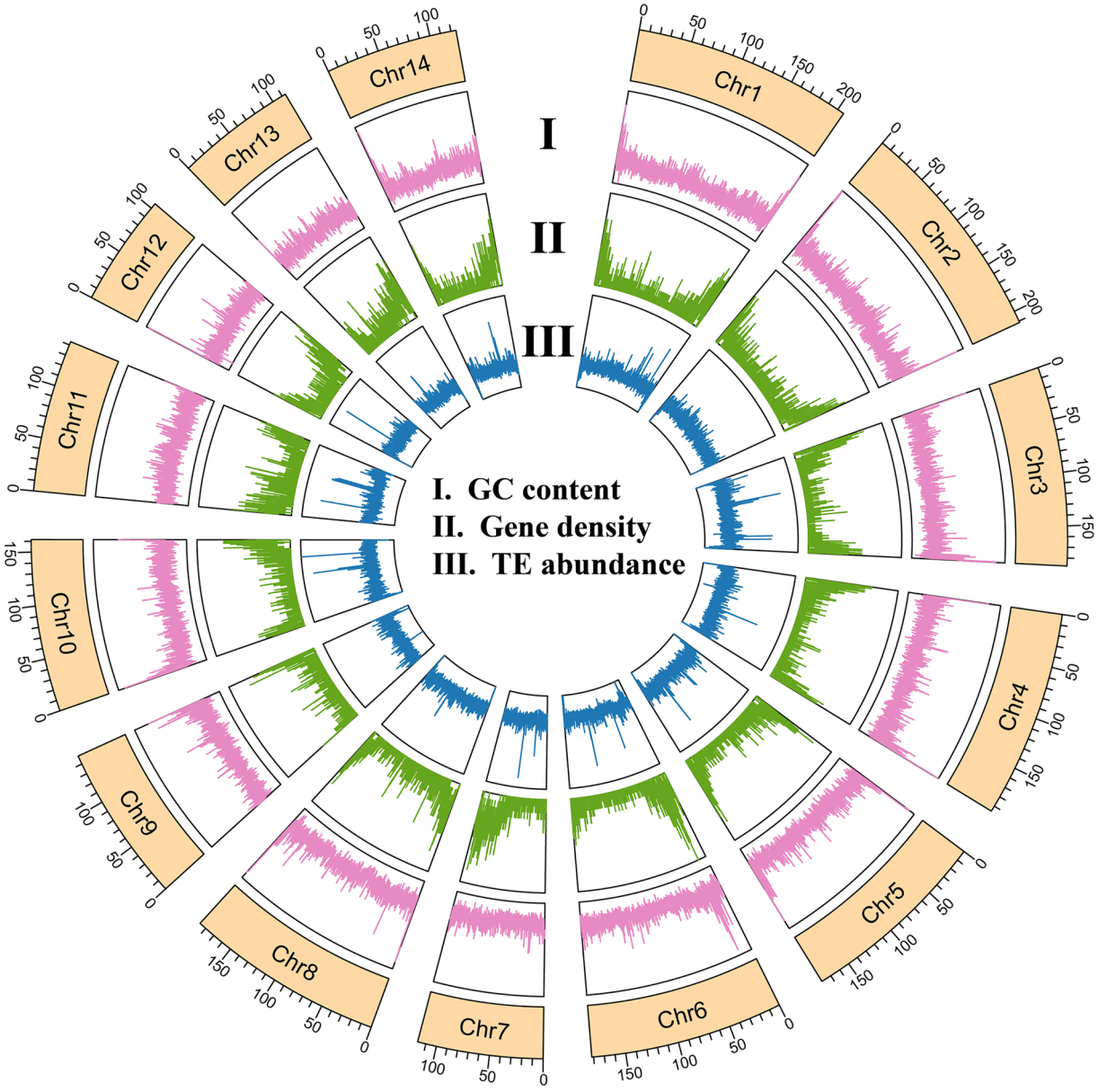

Fig. 4 Overview of genomic features of *Colocasia esculenta*. (I) GC content, (II) gene density, and (III) abundance of transposable elements.

	Number	Percent (%)
Total	28,253	NA
InterPro	21,336	75.52
GO	20,017	70.85
KEGG_ALL	26,339	93.23
KEGG_KO	9,591	33.95
Swissprot	18,022	63.79
TrEMBL	26,646	94.31
NR	26,717	94.56
Annotated	26,881	95.14
Unannotated	1,372	4.86

Table 7. Information on the number and ratio of functionally annotated genes based on different sequence databases.

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		**68**	129.18	8,784	0.000376
tRNA		**1,392**	74.68	103,951	0.004448
rRNA		**2,842**			
	rRNA	1,421	2,678.54	3,806,201	0.162881
	18S	418	1,952.48	816,138	0.034925
	28S	421	6,943.55	2,923,234	0.125095
	5S	582	114.83	66,829	0.00286
snRNA		**452**			
	snRNA	226	122.13	27,601	0.001181
	CD-box	100	111.89	11,189	0.000479
	HACA-box	3	146.00	438	0.000019
	splicing	123	129.87	15,974	0.000684
	scaRNA	0	0.00	0	0.00

Table 8. The number and length of the non-coding RNAs annotated in the assembled genome of Colocasia esculenta.

	Genome		Gene	
Type	**Count**	**Ratio (%)**	**Count**	**Ratio (%)**
Complete BUSCOs (C)	1,564	96.9	1,435	88.9
Complete and single-copy BUSCOs (S)	1,497	92.8	1,153	71.4
Complete and duplicated BUSCOs (D)	67	4.2	282	17.5
Fragmented BUSCOs (F)	34	2.1	99	6.1
Missing BUSCOs (M)	16	1.0	80	5.0
Total BUSCO groups searched	1,614	100.0	1,614	100.0
Complete HOGs			8,119	88.69
Complete and single HOGs			5,464	59.68
Complete and duplicated HOGs			2,655	29.0
Missing HOGs			1,036	11.32
Total HOGs searched			9,155	100.0

Table 9. Result of the BUSCO assessment of *Colocasia esculenta* genome and gene annotation.

genome. This achieved a 99.96% mapping rate and 99.99% coverage of HiFi reads. Taken together, the results from BUSCO, LAI, KAT, and Merqury analyses indicate that the genome assembly is of high quality, showing excellent completeness and base-level accuracy.

The completeness of the annotated protein-coding gene set in *C. esculenta* was further assessed using BUSCO with the embryophyta_odb10 database and OMArk v 0.3.0[53] against a reference set of 9155 hierarchical orthologous groups (HOGs) derived from Liliopsida. BUSCO analysis indicated that 88.9% of the conserved orthologs were present in the gene set (71.4% as single-copy BUSCOs), which is consistent with the OMArk completeness estimate of 88.7% (59.7% as single-copy HOGs) (Table 9).

Data availability

The raw sequencing data generated in this study have been deposited in the in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA1305327. The chromosome-level genome assembly has been deposited at GenBank under the accession JBQGWC000000000. The annotation files, coding sequences (CDS), and protein sequences are available on Figshare under https://doi.org/10.6084/m9.figshare. 29917034.

Code availability

All analyses followed the manuals of publicly available tools, with no custom code used.

Received: 29 August 2025; Accepted: 27 January 2026;
Published online: 09 April 2026

References

1. Food and Agriculture Organization. FAO crop production statistics https://www.fao.org/faostat/en/#data/QCL/visualize (2024).
2. Miyasaka, S. C. *et al.* Genetic Diversity in Taro (*Colocasia esculenta*) (pp. 191–215). Springer, Cham. https://doi.org/10.1007/978-3-319-96454-6_7 (2019).
3. Chaïr, H. *et al.* Genetic diversification and dispersal of taro (*Colocasia esculenta* (L.) Schott). *PLoS One* **6**, e0157712, https://doi.org/10.1371/journal.pone.0157712 (2016).
4. Aditika, Kapoor, B., Singh, S. & Kumar, P. Taro (*Colocasia esculenta*): Zero wastage orphan food crop for food and nutritional security. *S. Afr. J. Bot.* **145**, 157–169, https://doi.org/10.1016/j.sajb.2021.08.014 (2022).

5. Yin, J. *et al*. A high-quality genome of taro (*Colocasia esculenta* (L.) Schott), one of the world's oldest crops. *Molecular ecology resources* **21**(1), 68–77, https://doi.org/10.1111/1755-0998.13239 (2021).
6. Li, H., Liu, L., Qiu, Z., He, F. & Dong, W. Complete mitochondrial genome assembly and comparative analysis of *Colocasia esculenta*. *BMC Plant Biology* **25**(1), 67, https://doi.org/10.1186/s12870-025-06082-z (2025).
7. Takaesu, R. *et al*. Sequencing of the chloroplast genome of Taimo, a paddy cultivar of Taro (*Colocasia esculenta* (L.) Schott) in the Ryukyu Archipelago. *Mitochondrial DNA*. Part B, Resources, 10(6), 508–512. https://doi.org/10.1080/23802359.2025.2505789 (2025).
8. Dong, W., He, F., Wei, S., Qiu, Z. & Chen, Q. Identification and characterization of SSR markers in taro (*Colocasia esculenta* (L.) Schott) by RAD sequencing. *Genet. Resour. Crop Evol.* **68**, 2897–2905, https://doi.org/10.1007/s10722-021-01162-z (2021).
9. Ferdaus, M. J., Chukwu-Munsen, E., Foguel, A. & da Silva, R. C. Taro roots: An under exploited root crop. *Nutrients* **15**, 3337, https://doi.org/10.3390/nu16131983 (2023).
10. Eid, J. *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138, https://doi.org/10.1126/science.1162986 (2009).
11. Lieberman-Aiden, E. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293, https://doi.org/10.1126/science.1181369 (2009).
12. 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).
13. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
14. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
15. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
16. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
17. 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).
18. Chen, N. S. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **4**, 4.10, https://doi.org/10.1002/0471250953.bi0410s05 (2004).
19. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
20. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
21. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–80, https://doi.org/10.1093/nar/27.2.573 (1999).
22. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
23. *China National Center For Bioinformation (CNCB)* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRA028667 (2024).
24. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
25. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515, https://doi.org/10.1038/nbt.1621 (2010).
26. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
27. Gao, Y. *et al*. A chromosome-level genome assembly of *Amorphophallus konjac* provides insights into konjac glucomannan biosynthesis. *Comput Struct Biotechnol J.* **20**, 1002–1011, https://doi.org/10.1016/j.csbj.2022.02.009 (2022).
28. Wang, Y. *et al*. Chromosome level genome assembly of colored calla lily (*Zantedeschia elliottiana*). *Sci Data* **10**, 605, https://doi.org/10.1038/s41597-023-02516-1 (2023).
29. Qian, Z. *et al*. The chromosome-level genome of a free-floating aquatic weed *Pistia stratiotes* provides insights into its rapid invasion. *Mol Ecol Resour.* **22**, 2732–2743, https://doi.org/10.1111/1755-0998.13653 (2022).
30. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, https://doi.org/10.1186/1471-2105-6-31 (2005).
31. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467, https://doi.org/10.1093/nar/gki458 (2005).
32. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–9, https://doi.org/10.1093/bioinformatics/bth315 (2004).
33. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinform.* **48**, 4.11.1–4.11.39, https://doi.org/10.1002/0471250953.bi0411s48 (2014).
34. Chen, C. *et al*. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol Plant.* **13**, 1194–1202, https://doi.org/10.1016/j.molp.2020.06.009 (2020).
35. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354, https://doi.org/10.1093/nar/gkaa977 (2021).
36. Deng, Y. Y. *et al*. Integrated NR Database in Protein Annotation System and Its Localization. *Comput. Eng.* **32**, 71–74 (2006).
37. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334, https://doi.org/10.1093/nar/gkaa1113 (2021).
38. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
39. 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, https://doi.org/10.1093/nar/gky1085 (2019).
40. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169, https://doi.org/10.1093/nar/gky092 (2017).
41. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
42. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
43. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
44. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441, https://doi.org/10.1093/nar/gkg006 (2003).
45. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/nar/gkg006 (2013).

46. *NCBI BioProject* https://identifiers.org/ncbi/insdc.sra:SRP608616 (2025).
47. *NCBI GenBank* http://identifiers.org/ncbi/insdc:JBQGWC000000000 (2025).
48. Sun, Y. *et al*. *Chromosome level genome assembly of taro (Colocasia esculenta) Figshare* https://doi.org/10.6084/m9.figshare.29917034 (2025).
49. Manni, M. *et al*. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
50. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a k-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 756–758, https://doi.org/10.1093/bioinformatics/btw663 (2017).
51. 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).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
53. Nevers, Y. *et al*. Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.* **43**, 124–133, https://doi.org/10.1038/s41587-024-02147-w (2025).

Acknowledgements

This research was supported by the China Agriculture Research System of MOF and MARA with grant agreement ID CARS-24.

Author contributions

Honglian Zhu, Xiaobo Ying and Gang Song conceived and designed the study. Honglian Zhu acquised the funding and revised the manuscript. Haining Wang, Ying Wang, Ziyan Qian, Yuping Liu, Yingnan Yang and Xuxiong Qin collected the samples. Zhixin Wang, Haining Wang and Gang Huang analyzed the data. Yalin Sun and Zhixin Wang wrote the draft manuscript. All authors contributed to this manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Y., G.S. or H.Z.