scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-Level Genome Assembly and Annotation of purple nutsedge (*Cyperus rotundus* Cyperaceae)

Wenlong Wen[1,2,4], Yijie Zhang[1,2,4], Yating Feng[1,2], Rou Xu[1,2], Tianyu Li[1,2], Ziwei Yu[1,2], Honggang Wang[1,2] ✉, Fei Chen[1,2] ✉ & Yinhua Chen[1,2,3] ✉

***Cyperus rotundus*, one of the most aggressive weeds worldwide, demonstrates significant resistance to conventional control strategies. This species causes yield losses ranging from 20% to 90% in both agricultural and horticultural crops, posing a substantial threat to agricultural ecosystems. In this study, we present a chromosome-level genome assembly and annotation of *C. rotundus*. The haplotypic assembly spans approximately 293 Mb in length, with a high contig N50 of 5.49 Mb, a GC content of 36%, and a BUSCO score of 93.4%. A total of 94.03% (275.55 Mb) of the contigs were successfully mapped onto 54 chromosomes. Repetitive sequences constitute 44.78% of the genome. We predicted 23,280 protein-coding genes, 97% (52,439) of which were functionally annotated. The reference genome will serve as a valuable resource for phylogenomic studies within the *Cyperaceae* family and support further research on *C. rotundus*.**

Background & Summary

Agricultural weeds compete with crops for essential resources such as light, nutrients, and water, resulting in reduced crop yields and significant economic, environmental, and ecological consequences[1,2]. It is estimalia, 11 billion USD in India, and 33 billion USD in the United States, contributing to a global loss of around 200 million tons in food production[2]. The Cyperaceae family comprises about 3,000 species, with approximately 300 classified as weeds, and about 42% of these belong to the genus *Cyperus*[3].Among these, *C.rotundus* stands out as the most aggressive weed worldwide, due to its widespread distribution and its ability to outcompete crops[4]. Furthermore, this species demonstrates resistance to conventional control methods, making it one of the most problematic weeds globally, responsible for yield losses ranging from 20% to 90% in both crops and horticultural plants[3,5]. Consequently, *C. rotundus* exerts severe impacts on agricultural ecosystems across various regions.

C. rotundus (Family: Cyperaceae) derives its genus name from the ancient Greek word "Cypeiros" and its species epithet, "*rotundus*", from the Latin term for "round," referencing its round-shaped tubers[6]. Native to South Asia, Africa, Central and Southern Europe, and tropical, subtropical, and temperate regions of Australia[7], *C. rotundus* is an upright, hairless, grass-like perennial herb characterized by fibrous roots and slender, scaly, creeping rhizomes (Fig. 1). This species is commonly found in temperate, tropical, and subtropical regions, including China, India, South Africa, Korea, Japan, Egypt, and Iran[8–12]. Its primary mode of reproduction is asexual, through underground tubers, rhizomes, and basal bulbs, although sexual reproduction occurs via seeds[13]. *C. rotundus* exhibits remarkable reproductive and survival capabilities, thriving in dryland crop fields such as sugarcane[14], maize, soybeans[15], cotton[16] and peanuts[17], where it significantly affects crop growth and quality. In warm climates, *C. rotundus* is particularly difficult to manage due to its perennial nature, rapid growth, and high tuber production, making it a highly invasive species[18]. Despite the economic and ecological significance of this noxious weed, its complete genome has not yet been sequenced, and its biological characteristics and adaptive

[1]State Key Laboratory for Tropical Crop Breeding, Sanya Institute of Breeding and Multiplication, Hainan University, Sanya, 572025, China. [2]School of Tropical Agriculture and Forestry, Hainan University, Haikou, 570228, China. [3]Hainan Provincial International Joint Research Center for Utilization of Tropical Biological Resources, Hainan University, Haikou, 570228, China. [4]These authors contributed equally: Wenlong Wen, Yijie Zhang. ✉e-mail: deru666@hainanu.edu.cn; feichen@hainanu.edu.cn; yhchen@hainanu.edu.cn

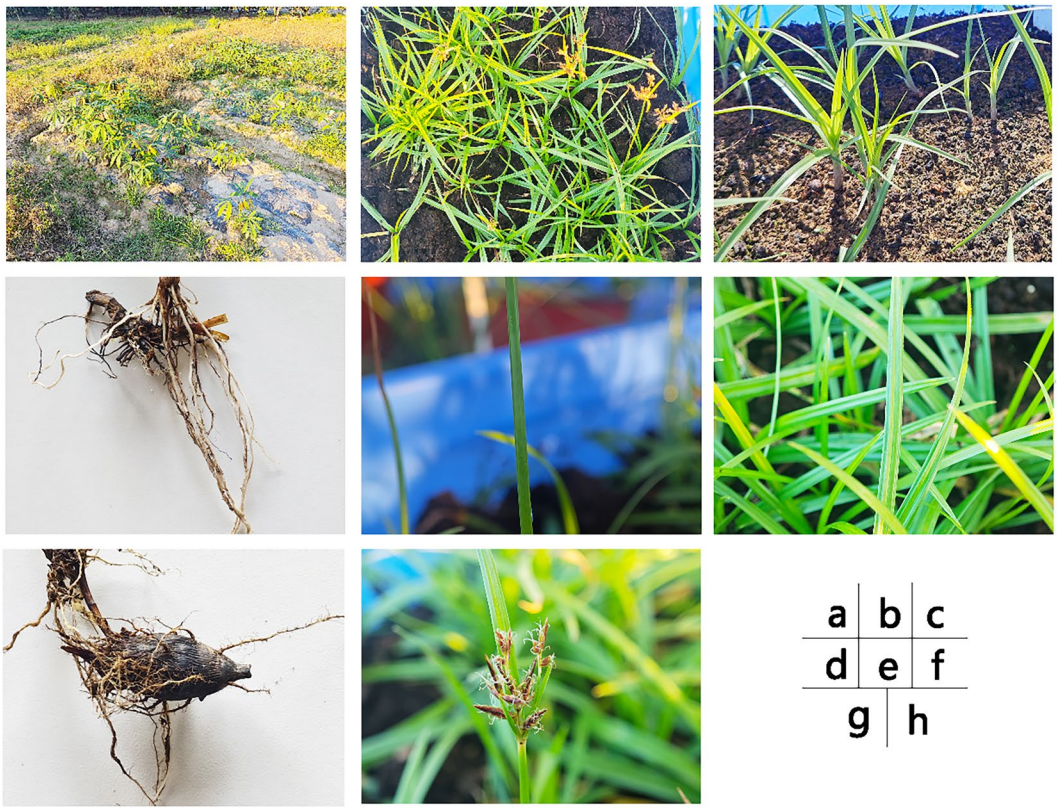

Fig. 1 Morphological characteristics of *C. rotundus*: (a) growth habit in cassava fields, (b) whole plant, (c) seedling, (d) root, (e) stem, (f) leaf, (g) tuber, (h) flower.

mechanisms remain inadequately understood. Therefore, the chromosome-level genome sequence of *C. rotundus* will help further elucidate its biological characteristics, adaptive mechanisms, and phylogenetic relationships.

In this study, we constructed and annotated a high-quality chromosome-level reference genome using integrated data (Fig. 2). The genome was initially assembled at the contig level using PacBio HiFi long reads and the hifiasm v0.18[19] tool. To achieve chromosome-level assembly, we employed Illumina Hi-C paired-end reads, processed through the HiCUP v0.9.2[20]. After masking repetitive sequences, we utilized three strategies for gene annotation with EVidenceModeler (EVM)[21]. These strategies included: homologous prediction based on closely related species, transcriptome prediction using Illumina paired-end RNA-seq short reads via the PASA[22] pipeline, and *de novo* prediction relying on genomic sequence features. Following gene function annotation for protein-coding genes and protein domains, we validated the results against relevant databases. The genomic features were then visualized by Circos v0.69.8[23] (Fig. 2). Additionally, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO)[24] analysis and evaluated genome mapping and coverage using Illumina paired-end short reads to assess the completeness and quality of both the genome assembly and annotation. These results demonstrate that the current genome assembly and annotation are both continuous and accurate. Thus, the present *C. rotundus* genomic resource will lay the foundation for further research on plants within this genus.

Methods

Plant material collection and preparation. In March 2024, mature and healthy *C. rotundus* plants were collected from the Danzhou campus of Hainan University in Danzhou City, Hainan Province (19°54′37.26″N, 109°31′48.67″E). The roots, stems, leaves, flowers, and tubers were harvested after being washed with deionized water. All tissues were immediately placed in liquid nitrogen and stored in a cryogenic freezer until further use.

DNA library construction and genome sequencing. DNA extraction was carried out using a modified cetyltrimethylammonium bromide (CTAB) method[25]. The quality and concentration of the extracted DNA were assessed through 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometry (Thermo Fisher Scientific, Wuhan), and Qubit 3.0 fluorometry (Life Technologies, Carlsbad, USA). For Illumina sequencing, high-quality DNA was defined as having an OD260/280 ratio of 1.6–1.8, no visible viscosity, a total amount $\geq$0.2 μg, a Qubit concentration $\geq$5 ng/μL, and intact bands on agarose gel electrophoresis. For ONT sequencing, high-quality DNA was defined as being clear and transparent, with no insoluble particles or viscosity, an Nc/Qc ratio of 0.95–1.5, A260/280 of 1.8–2.0, A260/230 $\geq$ 1.5, no degradation or only slight degradation, and a total amount $\geq$5 μg.After obtaining high-quality and purified genomic DNA, a SMRT cell sequencing library containing approximately 15–20 kb fragments was constructed and sequenced on the DNBSEQ-T7 platform. For PacBio

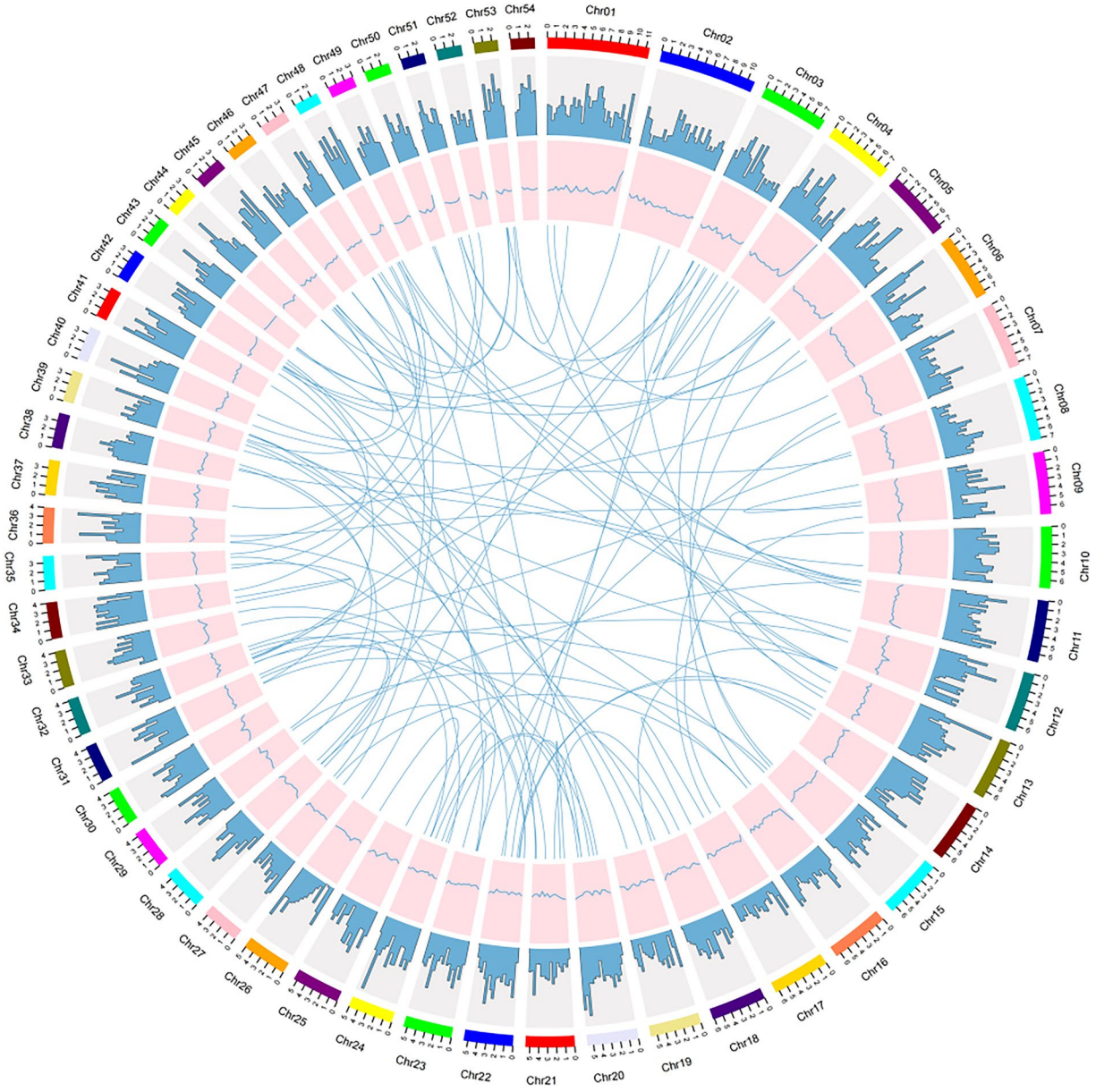

Fig. 2 Chromosome-scale genome assembly map of *C. rotundus*. From outermost to innermost, the Circos plot represents the following: (1) the length of 54 pseudo-chromosomes at the Mb scale; (2) gene density per Mb; (3) GC content per Mb; and (4) center: collinearity within *C. rotundus*.

HiFi sequencing, circular consensus sequencing (CCS) reads were generated using the ccs tool[26] in SMRT Link with parameters --min-passes 3 --min-length 10 --min-rq 0.99 to ensure high accuracy and remove low-quality data. Raw sequencing reads were processed to remove adapter sequences, low-quality reads (Q score < 20), and short fragments (<1 kb). A total of 12 Gb of clean data were generated, covering approximately 40.95 × of the genome (Table 1).

For Oxford Nanopore sequencing, the SQK-LSK110 ligation kit and standard protocol were used to prepare the library. The purified library was loaded onto an initialized R9.4 Spot-On flow cell and sequenced using the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). Basecalling was conducted using Guppy v6.3.8[27], during which reads with an average Q score below 7 were automatically filtered.Raw reads were filtered to remove adapter sequences, short reads (<1 kb), and low-quality reads (Q score <20), resulting in 12 Gb of clean data, which provided approximately 40.95 × genome coverage.Concurrently, an Illumina second-generation library (Illumina, San Diego, USA) was constructed with an insert fragment size of 350 bp. A total of 12 Gb of clean data, corresponding to approximately 40.95× genome coverage, was obtained after filtering with fastp v0.23.4[28], which removed low-quality reads, adapter sequences, and polyG tails using default settings.This short-read data was then utilized for genomic analysis.

For chromosomal scaffolding, a Hi-C library was prepared from *C. rotundus* tissues and sequenced using the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA). Raw sequencing data were processed using

Platform	Type	Sample	Molecule	Total clean data	Coverage	Usage	SRA accession number
Illumina NovaSeq	PE	Leaf	DNA	12 Gb	40.95X	correction	SRR34082906
ONT	UL	Leaf	DNA	12 Gb	40.95X	de novo assembly	SRR34082905
PacBio HiFi	CCS	Leaf	DNA	12 Gb	40.95X	de novo assembly	SRR34082903
Illumina Hi-C	PE	Leaf	DNA	23 Gb	78.49X	chromosome-level assembly	SRR34082904
Illumina NovaSeq	PE	Root	RNA	6.00 Gb	20.47X	gene structure annotation	SRR34082902
Illumina NovaSeq	PE	Stem	RNA	6.00 Gb	20.47X	gene structure annotation	SRR34082901
Illumina NovaSeq	PE	Leaf	RNA	6.00 Gb	20.47X	gene structure annotation	SRR34082899
Illumina NovaSeq	PE	Flower	RNA	6.00 Gb	20.47X	gene structure annotation	SRR34082898
Illumina NovaSeq	PE	Tuber	RNA	6.00 Gb	20.47X	gene structure annotation	SRR34082900

Table 1. Genomic Assembly and Annotation Sequencing Data Statistics of *C. rotundus*.

fastp[28] to remove adapter sequences, low-quality reads (Q < 20), and reads with ambiguous bases. After initial filtering, 23 Gb of clean Hi-C data were obtained, achieving approximately 78.49 × coverage. Hi-C data preprocessing was performed using HiCUP[20]. First, the reference genome was digested in silico using: hicup_digester–genome genome–re1 ^GATC,MboI genome.fasta, simulating MboI restriction enzyme cleavage. The paired-end Hi-C reads were then aligned to the draft genome assembly using Bowtie2 within the HiCUP pipeline via: hicup–bowtie2 bowtie2–digest Digest_genome*.txt–format Sanger–index genome–outdir $PWD–threads 4 $HiC_data_dir/*.fq.gz. Invalid read pairs-including self-ligated, re-ligated, circularized, and uninformative pairs-were automatically filtered. Valid paired-end reads were extracted using: samtools fastq −1 hicup.sam.tmp_R1.fastq −2 hicup.sam.tmp_R2.fastq hicup.sam, and all clean pairs were concatenated into final read files for scaffolding. Sequencing was performed by Wuhan Baita Gene Technology Co., Ltd. (Wuhan, China) (Table 1).

RNA library construction and transcriptome sequencing. Total RNA was extracted from the roots, stems, leaves, flowers, and tubers following the standard TRIzol protocol (Invitrogen, USA)[29]. Approximately 100 mg of tissue was ground into a powder with liquid nitrogen, followed by the addition of 1000 μL TRIzol into a 2.0 mL tube. The solution was incubated for about 5 minutes, then 200 μL of chloroform was added. The mixture was vigorously shaken for 30 seconds and allowed to stand for 3 minutes. After centrifugation at 12,000 rpm for 15 minutes at 4 °C, the upper aqueous phase was collected. This phase was then extracted with 500 μL isopropanol into a 1.5 mL tube and gently inverted to mix. After standing for approximately 10 minutes, the mixture was centrifuged at 12,000 rpm for 10 minutes. The supernatant was discarded, the pellet was washed twice with 75% ethanol, and the final pellet was dissolved in 50 μL of DNase- and RNase-free water for further analysis.

For Illumina paired-end sequencing, mRNA was reverse-transcribed into cDNA, and four libraries with an insert size of 350 bp were constructed following the manufacturer's instructions using the TruSeq RNA library preparation kit (Illumina, USA). Whole-genome shotgun sequencing was performed on the Novaseq 6000 platform using the PE 150 program. Raw sequencing data were processed using fastp[28] with parameters "--detect_adapter_for_pe -- qualified_quality_phred 5 --unqualified_percent_limit 50 --n_base_limit 5 --dedup" to remove adapter sequences, reads with ≥50% low-quality bases (Phred score ≤ 5),reads containing >5% ambiguous bases (N), and putative PCR duplicates. After filtering, a total of 40 Gb of clean data were obtained from the RNA-seq libraries (Table 1).

Genome assembly and Hi-C scaffolding. Initially, we estimated the genome size and heterozygosity of *C. rotundus* using K-mer analysis with Jellyfish v1.1.10[30] (parameter "-m 21") and GenomeScope v2.0[31] (parameter "k = 21") using clean Illumina short-read data. The K-mer analysis indicated that the genome size of *C. rotundus* is approximately 0.29 Gb (Fig. S1). To assemble third-generation long-read sequencing data into genomic sequences, tools such as Flye[32], Canu[33], and Hifiasm[19] are widely used. Genome assembly was performed using a hybrid approach combining PacBio HiFi and Oxford Nanopore (ONT) long reads. The assembly pipeline utilized Hifiasm v0.18[19], with parameters optimized for integrating ultra-long ONT data into the HiFi assembly. Specifically, the following command was used for hybrid genome assembly: nohup /software/hifiasm/hifiasm-0.18/hifiasm/hifiasm -t 4 -l 3 --ul ../ONT.fastq.gz -o HIFI_ONT_assembly ../HIFI_data.fastq 2 > error.txt &. After assembly completion, the primary contig fasta file was extracted from the GFA output using the command: awk '/^S/{print" > "$2;print$3}' HIFI_ONT_assembly.bp.p_ctg.gfa > genome.fasta. This approach enabled us to leverage the high base accuracy of PacBio HiFi reads and the ultra-long coverage of ONT reads, producing a draft genome suitable for subsequent scaffolding and annotation. To correct errors in the initial assembly, Illumina-derived short reads were employed for correction using Pilon v1.23[34]. The final *C. rotundus* genome assembly comprises 201 scaffolds, each corresponding to a chromosome (including ChrUN).This number reflects the scaffold-level assembly. However, since the assembly has not reached telomere-to-telomere (T2T) completeness, gaps still exist within certain chromosomes. These gaps divide chromosomes into multiple contiguous sequences (contigs), resulting in a total of 237 contigs (Table 2).

The assembled genome was further evaluated using the Benchmarking Universal Single-Copy Orthologs BUSCO v5.4.3[24] method with the embryophyta_odb10 dataset to assess the *C. rotundus* genome. The results indicated that 93.4% of BUSCO genes were successfully detected in the genome assembly, including 1,460 single-copy genes, 46 duplicated genes, 16 fragmented BUSCO genes, and 92 missing genes (Fig. 3).

Feature	Metric
Hifiasm-derived contigs	
Number of contigs	237
Total length of contigs	293,003,325 bp
Longest contig	8,964,331 bp
contig N50	4,270,926 bp (contig number = 26)
Contig N60	3,904,697 bp (contig number = 33)
Contig N70	3,195,727 bp (contig number = 42)
Contig N80	2,739,306 bp (contig number = 51)
Contig N90	1,743,940 bp (contig number = 63)
Hi-C scaffolded assembly	
Number of scaffolds	201
Total length of scaffolds	293,021,325 bp
Longest scaffold	11,392,672 bp
Scaffold N50	5,491,733 bp (scaffold number = 21)
Scaffold N60	5,048,567 bp (scaffold number = 26)
Scaffold N70	4,096,823 bp (scaffold number = 33)
Scaffold N80	3,536,775 bp (scaffold number = 41)
Scaffold N90	2,833,643 bp (scaffold number = 50)
GC content	36%

Table 2. Statistics of the *C. rotundus* genome assembly.

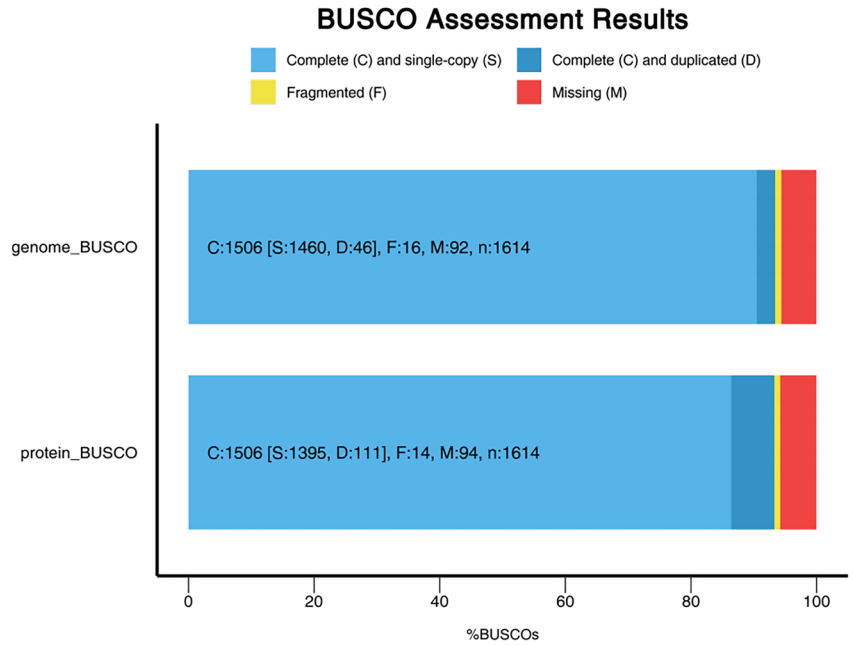

Fig. 3 Quality assessment of genome using embryophyta_odb10 database showed genome busco 93.4% and protein busco 93.3%. C: the number of complete genes, S: the number of complete and single-copy genes, D: the number of complete and duplicated genes, F: the number of incomplete genes, M: the number of missing genes.

To achieve a chromosome-level genome, the 237 contigs from the draft assembly were anchored onto 54 chromosomes using Haphic v1.0.6[35,36]. Hi-C reads were then aligned to the modified genome, and erroneous links, order, and orientation were manually corrected using Juicer v1.5[37]. After processing, Haphic was rerun. Ultimately, the Hi-C scaffolding resulted in chromosome-length scaffolds, producing 54 chromosomes (Fig. 4). The genome was purified of haplotypic duplications during the hybrid assembly process using HiFi and ONT reads, followed by chromosome-level scaffolding. Haplotypes were distinguished using synteny analysis with MUMmer software and the Hi-C heatmap, and the best-quality sequences from each haplotype were selected as the final genome assembly. The assumption of 54 chromosomes was based on our cytogenetic (karyotyping) analysis, which revealed that the organism has 162 chromosomes in total (Fig. S2). Furthermore, ploidy evaluation using the Smudgeplot[38] tool confirmed that *C. rotundus* is a triploid species (Fig. S3), and based on this, the

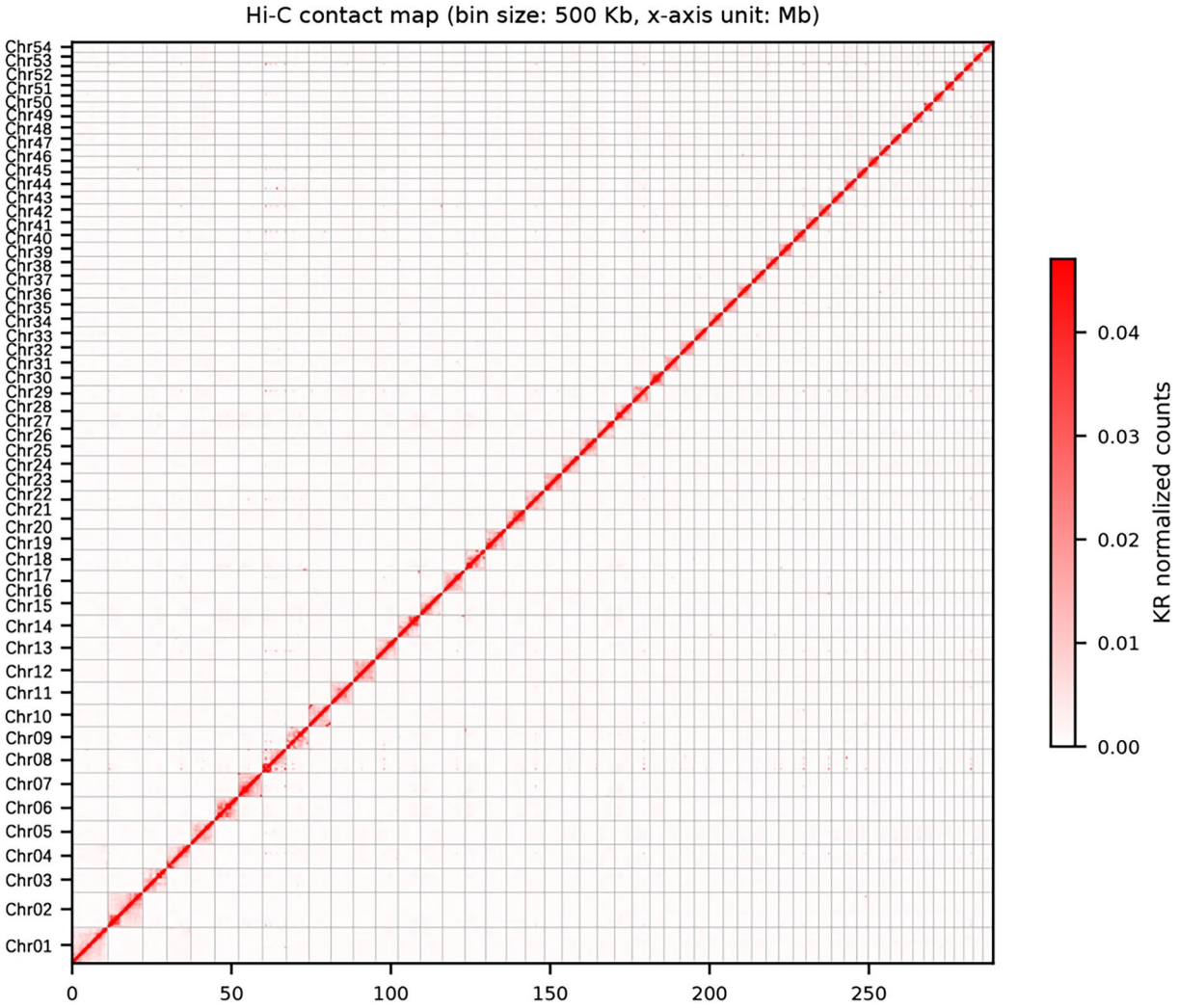

Fig. 4 The heatmap represents 54 chromosomes of the *C. rotundus* genome.

number of haploid chromosomes was inferred to be 54. Notably, haploid chromosome numbers around 54 have also been reported in other species of the Cyperus genus, and the presence of peaks at 18, 36, and 54 in haploid counts has been interpreted as potential evidence of polyploid series in this group[39].This value was used as the input parameter for HapHiC.

Genome annotation and functional prediction. *Repeat Sequence Annotation.* To identify repetitive sequences in the *C. rotundus* genome, we employed both homology-based methods and *de novo* prediction strategies.

Homology-based analysis. We first used RepeatMasker v1.323[40] and the Repbase TE library[41] to identify known transposable elements (TEs) in the *C. rotundus* genome.

De novo prediction. Next, we constructed a *de novo* repeat library for the genome using RepeatModeler open-1.0.8[42], which integrates two core *de novo* repeat detection tools-RECON v1.08[43] and RepeatScout v1.0.5[44] to identify and optimize dispersed repeat regions. To further explore long terminal repeat (LTR) retrotransposons, we conducted dedicated *de novo* searches using LTR_FINDER v1.0.7[45], LTR_harvest v1.5.11[46], and LTR_retriever v2.7[47]. Additionally, we employed Tandem Repeat Finder (TRF)[48] to identify tandem repeat sequences and MISA v1.0[49] to identify simple sequence repeats (SSRs).

Finally, we merged the repeat libraries obtained from the above methods and used RepeatMasker to annotate the repetitive content in the genome. The results indicated that 44.78% of the assemblies consisted of repetitive sequences (Table 3). The four main types of repetitive sequences were long terminal repeat sequences (LTR) (14.38% of the genome size), simple repeats (3.12%), DNA elements (2.46%), and unclassified elements (23.86%) (Table 3).

Type	Size (bp)	Percent of genome (%)
LINE	1,857,106	0.63
LTR	42,139,435	14.38
DNA	7,209,211	2.46
Unclassifed	69,908,764	23.86
Simple repeats	9,149,175	3.12
Low complexity	978,631	0.33
Total	131,242,322	44.78

Table 3. Statistics of repeat elements in the genome of *C. rotundus*.

Class	Type	Numbers	Average length (bp)	Total length (bp)	Percentage in genome (%)
miRNA		23,280	4,849.2321305	112,890,124	0.00589
tRNA		1,187	74	87,945	0.0003001
rRNA	18 S	314	1,810.99	568,653	0.00564
	28 S	316	4,018.14	1,269,734	0.00121
	5.8 S	320	153.93	49,260	0.00045
	5 S	129	118.24	15,254	0.01217
	8 S	129	113.81	14,681	0.01217
snRNA		120	98.26	11,792	0.00406
snoRNA		1,135	105.49	119,735	0.04122

Table 4. Non-coding RNAs in the *C. rotundus* assembly.

Non-coding RNA annotation. We used Rfam v14.0[50] to predict ribosomal RNA (rRNA), small nuclear RNA (snRNA), and microRNA (miRNA) by comparing the *C. rotundus* genome with known non-coding RNA libraries. The tRNAscan-SE v1.3.1[51] algorithm with default parameters was used to identify tRNA-related genes. In total, 3,650 non-coding RNAs (ncRNAs) were identified in the *C. rotundus* genome, including 1,187 tRNAs, 1,208 rRNAs, 23,280 miRNAs, 120 snRNAs, and 1,135 snoRNAs (Table 4).

Gene structure prediction. We used three strategies to predict the gene structure of the repeat-masked *C. rotundus* genome: initial gene prediction, homology-based gene prediction, and RNA-Seq guided gene prediction. Prior to gene prediction, the assembled *C. rotundus* genome was subjected to both hard and soft repeat masking using RepeatMasker[40], in order to improve annotation accuracy. Hard masking was applied to eliminate potential false gene predictions in highly repetitive regions, while soft masking was retained to preserve sequence context for downstream analyses that are sensitive to repeat content.

For initial gene prediction, we used Augustus v3.3.3[52]. Each gene prediction model was trained with a set of high-quality proteins generated from the RNA-Seq dataset. For homology-based gene prediction, we employed Maker v2.31.10[53], which was used to align protein and transcript sequences to our genome assembly and predict the coding genes.

For RNA-Seq guided gene prediction, we first used Hisat2 v2.0.0[54] to align the cleaned RNA-Seq reads to the genome. We then used Trinity v2.3.2[55], Transdecoder v2.01 (github.com/TransDecoder/TransDecoder), and Maker[54] to construct the gene structures.

Finally, EVidenceModeler (EVM) v1.1.1[21] was employed to integrate the predictions from the three methods and generate the final gene models. The output consisted of consistent, non-overlapping sequence assemblies to define the gene structure. In total, 23,280 protein-coding genes were predicted in the *C. rotundus* genome, with an average gene length of 2,280 bp (Table 5).

Gene function annotation. Gene functions were inferred according to the best match of the alignments to the National Center for Biotechnology Information (NCBI) Non-Redundant (NR)[56], TrEMBL[57], KOG[58] and Swiss-Prot protein databases[59] using BLASTP v2.6.0[60] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[61] with an E-value threshold of 1E-5. The protein domains were annotated using PfamScan v1.6[62] based on PFAM databse[63] and InterPro protein database[64]. Gene Ontology (GO) IDs for each gene were obtained from Blast2GO[65]. In total, approximately 97% of the predicted protein-coding genes of *C. rotundus* genome could be functionally annotated with known genes, conserved domains, and Gene Ontology terms(Table 6).

Data Records

Raw sequencing data have been deposited in the NCBI Sequence Read Archive (SRA)(Table 1) under BioProject accession number PRJNA1280253 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1280253), including transcriptomic, PacBio HiFi, Hi-C, Oxford Nanopore Technologies (ONT), and Illumina sequencing datasets (accessions SRR34082898–SRR34082906)[4,66–74].

Version	Gene number	Total Transcripts length	GC content(%)
Augustus Transcripts	28,500	32,394,514	48.07
Maker Transcripts	20,465	23,095,557	47.37
RNA Iso based Transcripts	17,148	43,880,069	42.87
Final Transcripts	23,280	53,096,603	43.08

Table 5. Statistics of gene structure prediction results.

Data base	Annotated number	Annotated ratio
GO	28,574	53%
KEGG	17,870	33%
KOG	30,313	56%
NR	51,802	96%
PFAM	46,045	86%
SWISS_PROT	42,768	79%
TREMBL	51,815	96%
Total	52,439	97%

Table 6. Statistics for the *C. rotundus* functionally annotated protein-coding genes.

The chromosome-level genome assembly of Cyperus rotundus is available at NCBI GenBank under the accession GCA_052426515.1 (https://www.ncbi.nlm.nih.gov/assembly/GCA_052426515.1)[75], linked to BioProject PRJNA1283543 and BioSample SAMN49699731.

The genome annotation files are accessible via Figshare (https://doi.org/10.6084/m9.figshare.29435915)[76].

Technical Validation

Genomic integrity, fragmentation, and possible loss rates were measured using BUSCO. The completeness of the protein sequences aligns with the genomic assessment results, indicating that the assembled genome has high integrity in terms of protein-coding genes. Among the 1,614 conserved core genes in the *Embryophyta* database, 1,506 (93.3%) were identified as complete BUSCO, and 16 (0.99%) as fragmented BUSCO, indicating that the assembled genome exhibits high integrity and validity, making it suitable for further analysis (Fig. 3).

Data availability

All sequencing data, genome assemblies, and annotations supporting this study are publicly available. Raw sequencing data have been deposited in the NCBI SRA under BioProject PRJNA1280253 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1280253). The genome assembly is available in GenBank under accession GCA_052426515.1[75], and annotation files are accessible from Figshare (https://doi.org/10.6084/m9.figshare.29435915)[76].

Code availability

All data processing commands and pipelines were executed according to the instructions and guidelines provided by the relevant bioinformatics software. No custom scripts or code were used in this study.

Received: 8 April 2025; Accepted: 8 October 2025;
Published online: 19 November 2025

References

1. Oerke, E.-C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2006).
2. Chauhan, B. S. Grand challenges in weed management. *Front. Agron.* **1**, 3 (2020).
3. Bendixen, L. E. & Nandihalli, U. B. Worldwide distribution of purple and yellow nutsedge (*Cyperus rotundus* and *C. esculentus*). *Weed Technol.* **1**, 61–65 (1987).
4. Holm, L.G., Plucknett, D.L., Pancho, J.V. & Herberger, J.P. *The World's Worst Weeds: Distribution and Biology* (University Press of Hawaii, 1977).
5. Peerzada, A. M. Biology, agricultural impact, and management of *Cyperus rotundus* L.: the world's most tenacious weed. *Acta Physiol. Plant.* **39**, 270 (2017).
6. Vernon, V.V. & Jason, A.F. Purple Nutsedge, *Cyperus rotundus* L. (Institute of Food and Agricultural Sciences, University of Florida, 2012).
7. Babiaka, S. B. *et al.* Natural products in *Cyperus rotundus* L. (Cyperaceae): an update of the chemistry and pharmacological activities. *RSC Adv.* **11**, 15060–15077 (2021).
8. Chang, K. S. *et al.* Contact and fumigant toxicity of *Cyperus rotundus* steam distillate constituents and related compounds to insecticide-susceptible and -resistant *Blattella germanica*. *J. Med. Entomol.* **49**, 631–639 (2012).
9. Aeganathan, R. *et al.* Anti-oxidant, antimicrobial evaluation and GC-MS analysis of *Cyperus rotundus* L. rhizomes chloroform fraction. *Am. J. Ethnomed.* **2**, 14–20 (2015).
10. Liu, X. C. *et al.* Chemical composition and insecticidal activity of the essential oil of *Cyperus rotundus* rhizomes against *Liposcelis bostrychophila* (Psocoptera: Liposcelididae). *J. Essent. Oil Bear. Pl.* **19**, 640–647 (2016).
11. Janaki, S. *et al.* Chemical composition and insecticidal efficacy of *Cyperus rotundus* essential oil against three stored product pests. *Int. Biodeterior. Biodegrad.* **133**, 93–98 (2018).

12. Sabir, M. N., Saour, K. Y. & Rachid, S. *In vitro* cytotoxic and antimicrobial effects of a novel peroxysesquiterpene glucoside from the rhizomes of *Cyperus rotundus* L (Cyperaceae). *Trop. J. Pharm. Res.* **19**, 331–339 (2020).
13. Thullen, R. J. & Keeley, P. E. Seed production and germination in *Cyperus esculentus* and *C. rotundus*. *Weed Sci.* **27**, 502–505 (1979).
14. Durigan, J. C. Effects of plant densities and management of purple nutsedge on sugarcane yield and effect of growth stages and main way of herbicides contact and absorption on the control of tubers. *J. Environ. Sci. Health B* **40**, 111–117 (2005).
15. Tuor, F. A. & Froud-Williams, R. J. Interaction between purple nutsedge, maize and soybean. *Int. J. Pest Manage.* **48**, 65–71 (2002).
16. Salgado, T. P., Alves, P. L. C. A. & Rossi, C. V. S. Effect of purple nutsedge (*Cyperus rotundus*) tuber density on the initial growth of cotton plants. *Planta Daninha* **20**, 405–411 (2002).
17. Du, L. *et al*. Density effect and economic threshold of purple nutsedge (*Cyperus rotundus* L.) in peanut (*Arachis hypogaea* L.). *Int. J. Plant Prod.* **13**, 309–316 (2019).
18. Santos, B. M. *et al*. Effects of shading on the growth of nutsedges (*Cyperus* spp.). *Weed Sci.* **45**, 670–673 (1997).
19. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
20. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
21. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
22. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
23. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**(9), 1639–1645 (2009).
24. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
25. Allen, G. C. *et al*. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
26. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
27. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
28. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
29. Rio, D. C. *et al*. Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harb. Protoc.* **2010**, pdb.prot5439 (2010).
30. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
31. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
32. Kolmogorov, M. *et al*. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
33. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
34. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
35. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
36. Zhang, X. *et al*. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
37. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
38. 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 (2020).
39. Roalson, E. H. A Synopsis of Chromosome Number Variation in the Cyperaceae. *Bot. Rev* **74**, 209–393 (2008).
40. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10.1–4.10.14 (2004).
41. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 1–6 (2015).
42. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
43. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
44. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
45. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
46. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 1–14 (2008).
47. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
48. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
49. Beier, S. *et al*. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
50. Griffiths-Jones, S. *et al*. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
51. 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 (1997).
52. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
53. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 1–14 (2011).
54. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
55. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
56. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
57. O'Donovan, C. *et al*. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3**, 275–284 (2002).
58. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 1–14 (2003).
59. UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699–2699 (2018).
60. Altschul, S. F. *et al*. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
61. Kanehisa, M. *et al*. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

62. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
63. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
64. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
65. Götz, S. *et al*. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082898 (2025).
67. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082899 (2025).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082900 (2025).
69. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082901 (2025).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082902 (2025).
71. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082903 (2025).
72. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082904 (2025).
73. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082905 (2025).
74. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34082906 (2025).
75. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052426515.1 (2025).
76. Wen, W. Annotation files for *Cyperus rotundus* genome assembly. *Figshare* https://doi.org/10.6084/m9.figshare.29435915 (2025).

Acknowledgements

The work was supported by the earmarked fund for Chinese Agriculture Research System (CARS-11-hncyh).

Author contributions

Y.H.C., F.C. and H.G.W. conceived and designed the study; Y.J.Z. and Y.T.F. collected the samples and extracted the genomic DNA for sequencing; W.L.W. and Y.J.Z. assembled the genome; R.X. and L.Z.X. performed Hi-C scaffolding and genome annotation; T.Y.L., Z.W.Y. performed technical validation; W.L.W. wrote the draft manuscript. W.L.W. and Y.J.Z. modified the manuscript. All authors have read, revised, and approved the final manuscript for submission.

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/s41597-025-06126-x.

Correspondence and requests for materials should be addressed to H.W., F.C. or Y.C.