scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the dwarf cattail *Typha minima*

Junshuai Du, Lei Huang & Xinwei Xu ✉

We report the first chromosome-level genome assembly of the critically endangered dwarf cattail, *Typha minima*, a wetland species of ecological and medicinal importance. Utilizing PacBio HiFi long-read sequencing and Hi-C scaffolding technologies, we generated a high-quality 324.66 Mb genome, anchored onto 30 pseudochromosomes. The assembly demonstrates exceptional continuity, with contig and scaffold N50 values of 10.84 Mb and 10.90 Mb, respectively, and a near-complete chromosomal anchoring rate of 99.65%. It exhibits outstanding completeness, as reflected by a BUSCO score of 99.2%, and contains 33.20% repetitive sequences. We annotated 34,541 protein-coding genes, with 96.42% receiving functional assignments. The assembly also includes annotations for non-coding RNAs, comprising 1,261 rRNAs, 230 miRNAs, and 467 tRNAs. Integrated orthology analysis identified 10,055 consensus orthologs across five functional databases. This high-quality genomic resource provides a foundation for advancing studies in evolutionary adaptation and conservation genomics of this endangered wetland plant.

Background & Summary

The genus *Typha* (cattails) comprises large emergent aquatic macrophytes that constitute a vital component of wetland ecosystems[1]. These plants play a crucial role in maintaining ecosystem functions by enhancing structural complexity, regulating biogeochemical cycles, and contributing to overall productivity[2–4]. Notably, *Typha* species effectively remove pathogenic microorganisms from water through root adsorption and the secretion of antimicrobial compounds[5], highlighting their ecological importance in maintaining wetland health. Beyond their ecological contributions, *Typha* species are valued for their medicinal properties. In traditional medicine, their dried pollen, known as Puhuang, is recognized for its hemostatic, stasis-resolving, and diuretic effects[6]. Given these multifaceted benefits, *Typha* species possess considerable applied potential and merit further scientific investigation.

Typha minima, characterized by a delicate growth habit, is native to temperate Eurasia[7]. It is currently listed as endangered in Switzerland and persists only in small, isolated populations in several other European countries[8]. Phylogenetic analyses have shown that *T. minima* and *T. elephantina* form a monophyletic clade, which is sister to the clade containing all other *Typha* species[9]. To date, genomes have been published for three *Typha* species: *T. latifolia*, *T. angustifolia* and *T. domingensis*[10–12]. Therefore, obtaining a high-quality genome of *T. minima* is crucial for advancing *Typha* phylogenomic, elucidating the genetic mechanisms underlying its endangerment, and developing effective conservation strategies.

In this study, we constructed a high-quality chromosome-level genome assembly for *T. minima* by integrating PacBio HiFi long-read sequencing and Hi-C chromatin interaction data. The final assembly spans 324.66 Mb with a scaffold N50 of 10.90 Mb, and 99.65% of the assembled sequences were successfully anchored onto 30 pseudochromosomes (Table 1, Supplementary Table S1). The genome exhibits high completeness, supported by a BUSCO completeness score of 99.20% (1,601 of 1,614 conserved genes) and a sequencing reads mapping rate of 98.19% (Table 1). Repetitive elements accounted for 33.20% (107.80 Mb) of the genome (Table 2). Among these, long terminal repeat (LTR) retrotransposons were the most abundant class (12.48%), predominantly represented by Gypsy (5.13%) and Copia (1.15%) families. Non-LTR retrotransposons and DNA transposons comprised 0.75% and 0.61%, respectively. Additionally, 11.94% of repeats remained unclassified and may include novel lineage-specific elements. A total of 34,541 protein-coding genes were annotated, of which 33,304

National Field Station of Freshwater Ecosystem of Liangzi Lake, College of Life Sciences, Wuhan University, Wuhan, P. R. China. ✉e-mail: xuxw@whu.edu.cn

Statistical feature	Corresponding value
Assembled genome size (bp)	324,661,732
Number of contigs	33
Number of scaffolds	36
Contig N50 (bp)	10,842,424
Scaffold N50 (bp)	10,900,986
Number of chromosomes	30
Genome sequences anchored to chromosomes (bp)	323,512,042
Anchoring rate	99.65%
GC content	37.91%
BUSCO complete genes(C)	1,601 (99.2%)
BUSCO single copy genes(S)	971 (60.2%)
BUSCO duplicated genes(D)	629 (39%)
BUSCO fragmented genes(F)	6 (0.4%)
BUSCO missing genes (M)	6 (0.4%)

Table 1. Genome assembly summary of *Typha minima*.

Type	Length (bp)	% in genome
DNA transposon	5,008,465	1.54%
LINE	2,436,784	0.75%
SINE	266	0.00%
LTR	20,630,798	18.83%
Satellite	188,970	0.06%
Other	255,404	0.08%
Unknown	38,776,748	11.94%
Total	67,297,435	33.20%

Table 2. Summary of repetitive sequences in the *Typha minima* genome.

Statistical feature	Corresponding value
Total gene number	34,541
Total transcript number	49,360
Mean gene length (bp)	4,949
Functionally annotated gene number	33,304

Table 3. Summary of the protein-coding gene annotation for the *Typha minima* genome.

(96.42%) were assigned functional descriptions (Table 3). This chromosome-level genome of *T. minima* serves as a valuable genomic resource for investigating the genetic basis of its endangered status and will support further studies on the evolution and phylogenetic relationships within the genus *Typha*.

Methods

Genome sequencing and assembly. Fresh leaves of *Typha minima* were collected from Kashgar, Xinjiang, China (39°14'15.2"N, 76°09'41.4"E; elevation 1,228 m). The samples were immediately frozen in liquid nitrogen and stored at −80 °C until DNA and RNA extraction. Genomic DNA was extracted using a modified CTAB protocol[13]. DNA purity was assessed with NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) to determine A260/A280 and A260/A230 ratios Potential degradation and RNA contamination were evaluated using pulsed-field gel electrophoresis (PFGE) and Qubit 3.0 fluorometry (Thermo Fisher Scientific, USA). For initial genome survey, we generated 65 Gb (~196 × coverage) of 150-bp paired-end Illumina NovaSeq. 6000 reads (Supplementary Table S1). Raw reads were quality-filtered by removing reads containing >3 N-bases, adapter contamination, or ≥ 20% of bases with Phred Q < 20. Contamination screening was performed by aligning 10,000 randomly selected reads against the NT database[14] using BLASTN v2.12.0[15]. K-mer analysis (k = 23) conducted with Jellyfish v1.1.11[16] and GenomeScope v1.0[17] estimated a genome size of 331.55 Mb (Fig. 1). The GenomeScope profile exhibited three peaks with an approximate ratio of 1:2:4 (Fig. 1), suggesting that *T. minima* is likely a tetraploid. The inference was further confirmed by Smudgeplot[18] analysis, which supported an allotetraploid genome structure (Fig. 2). For high-quality genome assembly, we prepared PacBio HiFi libraries using the SMRTbell Express 2.0 Kit, with library quality assessed by capillary electrophoresis (Fragment Analyzer 5400, Agilent). High-fidelity reads were processed via the CCS module in SMRT Link v11.0, yielding 48.66 Gb of HiFi data (~147 × coverage; Supplementary Table S1). *De novo* assembly was performed using Hifiasm v0.19.5[19],

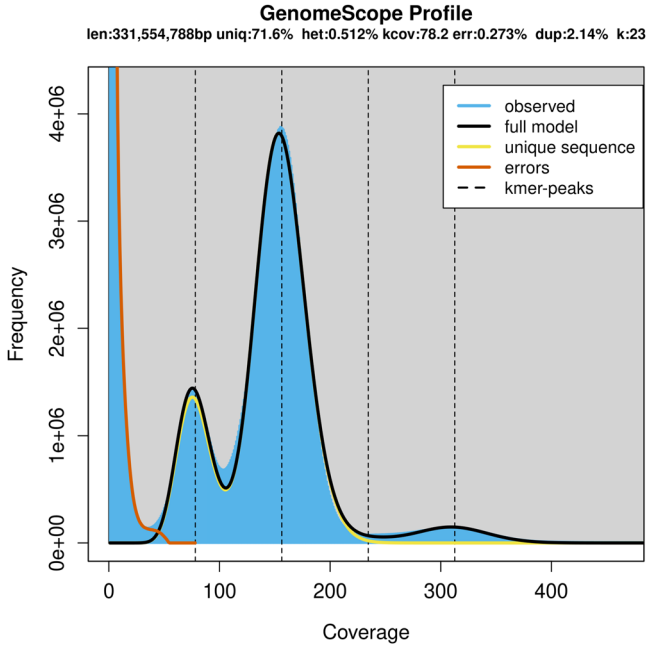

Fig. 1 Genome survey of *Typha minima* using k-mer distribution analysis (k = 23).

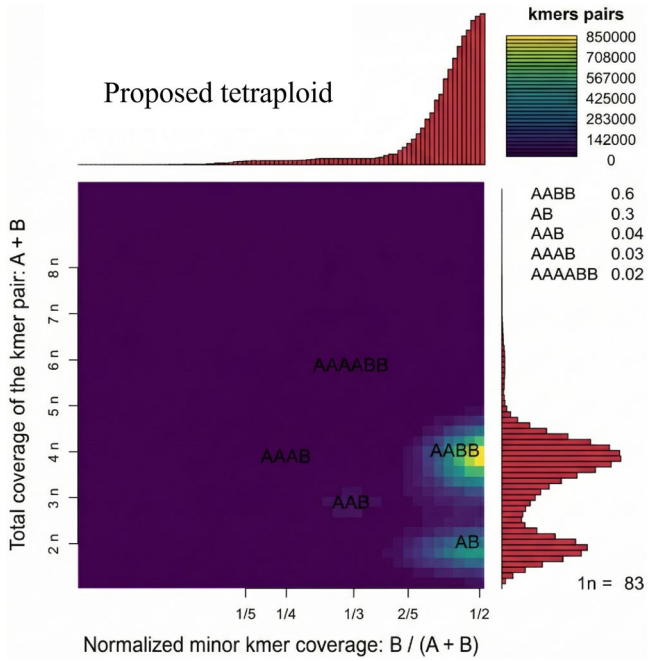

Fig. 2 The result of smudgeplot analysis for *Typha minima*.

followed by haplotype purging with purge_dups v1.2.5[20] to remove heterozygous redundancies, resulting in a final genome assembly of 324.66 Mb.

Following assembly, we performed comprehensive quality assessment through the analysis of 10-kb non-overlapping genomic windows for GC content and mean coverage depth. This dual-parameter analysis facilitated both base composition profiling and the detection of potential exogenous contamination. The absence of discrete clusters in bivariate GC-depth distributions (Fig. 3) confirmed a contamination-free assembly. For genome annotation, we implemented an integrated pipeline based on the embryophyta_odb10 single-copy ortholog set[21], consisting of three main steps: initial screening using TBLASTN[22], gene prediction with AUGUSTUS v3.5[23], and domain validation using HMMER v3.3.2[24]. The quality of the genome assembly was evaluated from two aspects: gene completeness and sequence contiguity. First, gene completeness was assessed using BUSCO v5.4.3[21] against the viridiplantae_odb10 database, yielding a completeness score of 99.20%. Subsequently, raw

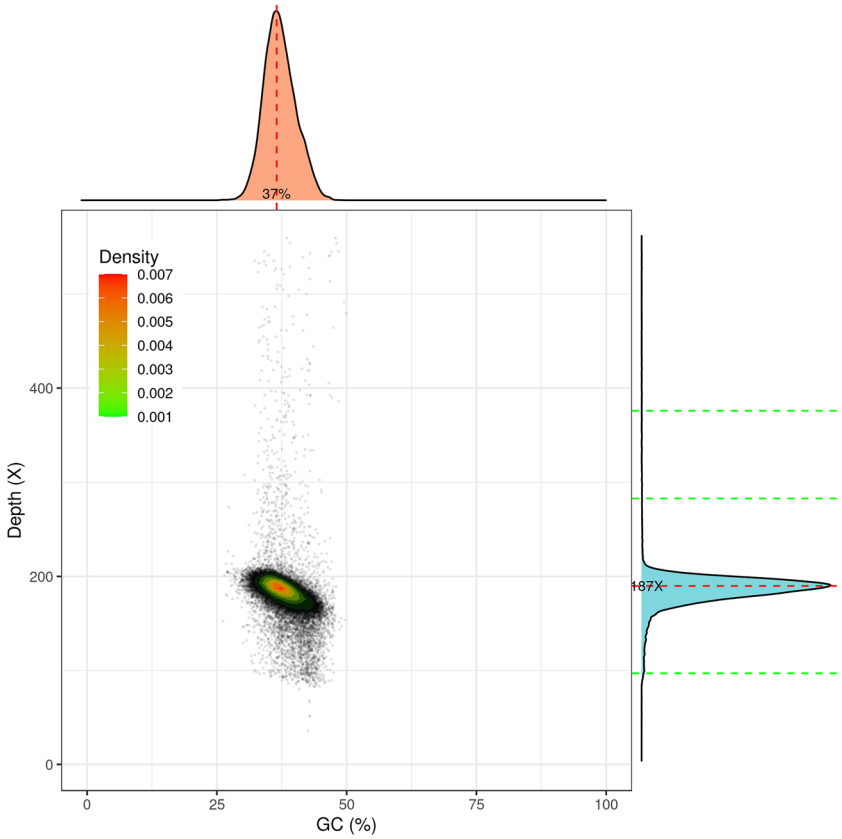

Fig. 3 GC content and depth distribution.

second-generation sequencing data were mapped back to the assembled genome using BWA v0.7.17[25] with default parameters, yielding a read mapping rate of 98.19% (Table 1). These results collectively demonstrate that the assembled genome possesses high completeness and accuracy.

Hi-C-Assisted genome assembly. Fresh leaf tissue was fixed with 1% formaldehyde for 15 min to preserve chromatin architecture, after which the crosslinking reaction was quenched using 0.125 M glycine. Hi-C libraries were constructed according to a standard protocol[26] and sequenced as 150-bp paired-end reads on an Illumina NovaSeq. 6000 platform, yielding 52.24 Gb of Hi-C data (~161 × coverage, Supplementary Table S1). Raw reads were quality-filtered, and 10,000 randomly selected clean reads were screened for contamination by aligning against the NT[14] database using BLASTN v2.12.0[15]. No exogenous contamination was detected (Supplementary Table S2). Valid chromatin interactions were aligned to the draft assembly using HiC-Pro v2.11.4[27]. Chromosome-level scaffolding was achieved through initial scaffolding with 3D-DNA v180922[28] and subsequently manual curation in Juicebox v1.11.08[29]. Genome-wide contact maps were generated using HiCExplorer v3.7.2[30] to visualize interaction intensities (Fig. 4). The final genome architecture was illustrated using Circos v0.69.9[31], highlighting key genomic features (Fig. 5).

Genome annotation. To optimize genome annotation, we performed transcriptome sequencing on root, stem, leaf, and fruit tissues of *T. minima* collected from Kashgar, Xinjiang. Total RNA was extracted from each tissue using the RNAprep Pure Plant Kit (Cat. No. DP441, Tiangen Biotech). RNA integrity, purity, and concentration were assessed using a NanoDrop spectrophotometer and an Agilent 2100 Bioanalyzer. Strand-specific libraries were constructed with the TruSeq mRNA-seq Kit and sequenced on the Illumina NovaSeq. 6000 platform (PE150). The four libraries generated 6.82 Gb, 7.32 Gb, 5.94 Gb, and 6.83 Gb of raw data, respectively, with Q30 base percentages exceeding 91% in all samples. To complement the short-read data, we performed PacBio full-length transcriptome sequencing. A SMRTbell library was prepared from leaf RNA using the Iso-Seq express 2.0 Kit and Kinnex Full-length RNA Kit, and sequenced on the PacBio Sequel II platform, generating 4.6 Gb of HiFi data. The resulting reads had an average length of 1,622 bp, a median length of 1,593 bp, an N50 of 1,776 bp, and a median base quality of Q35.

Repetitive elements were annotated through a multi-step pipeline. Initially, Miniature Inverted-repeat Transposable Elements (MITEs) were identified using MITE-Hunter v1.0[32] with default parameters. This was followed by LTR retrotransposon prediction via LTRharvest[33] and LTR_Finder v1.0.7[34], with results integrated using LTR_retriever v2.9[35]. An initial screening against the RepBase v20170127 database[36] was conducted with RepeatMasker v4.1.1[37], complemented by *de novo* prediction on masked sequences using RepeatModeler v2.0[38]. This comprehensive analysis identified 67.3 Mb repetitive sequences, accounting for 33.2% of the genome.

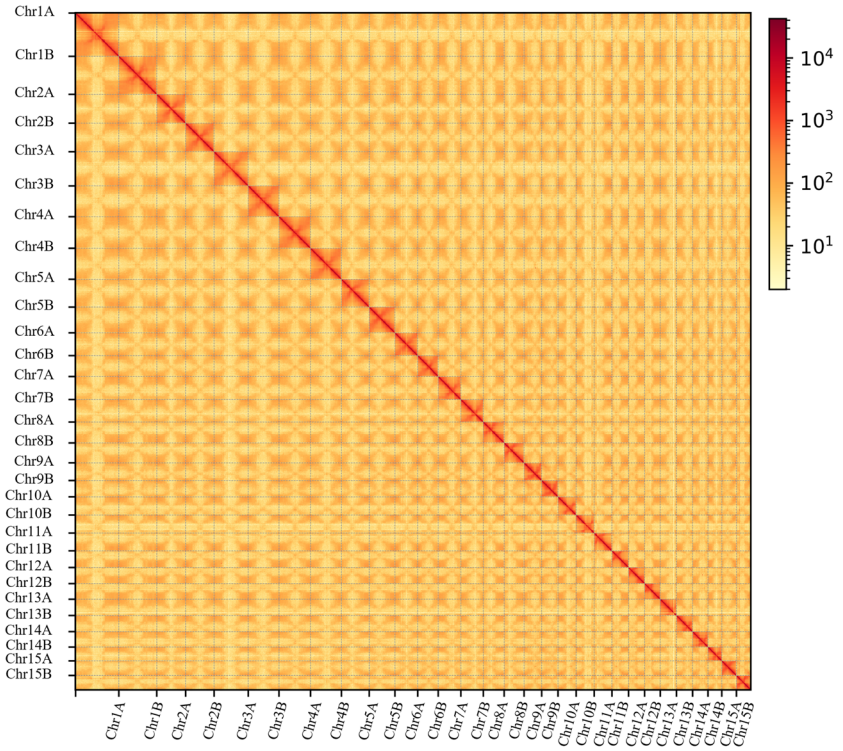

Fig. 4 Hi-C heatmap for the genome assembly of *Typha minima*.

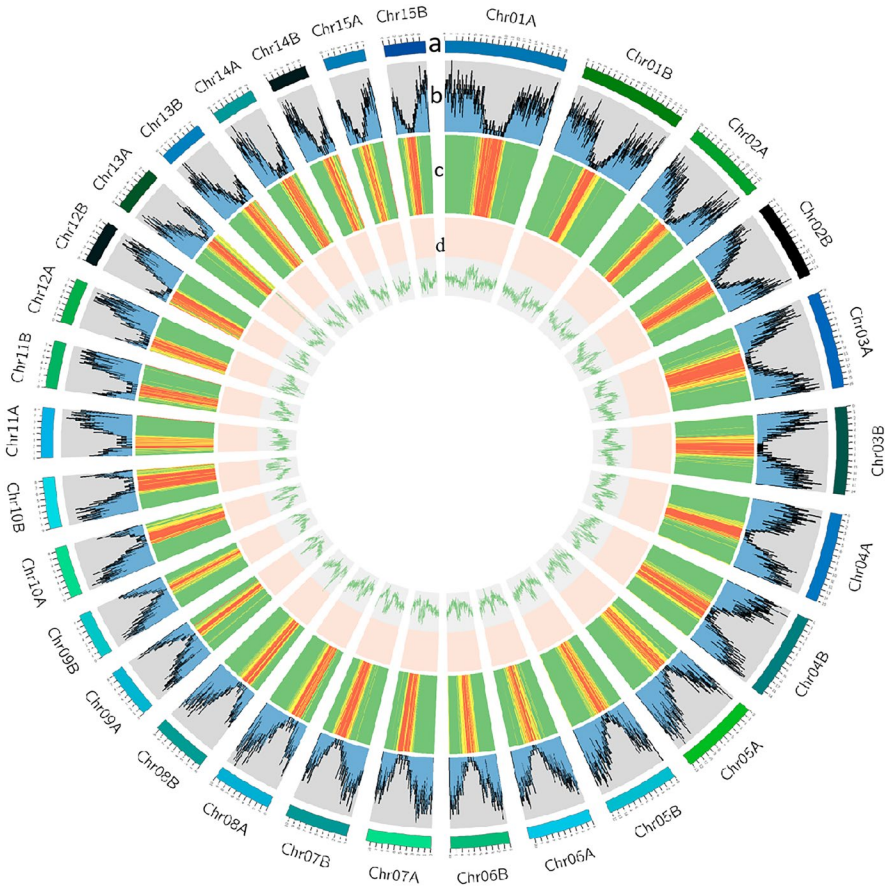

Fig. 5 Genomic features of *Typha minima*. The features are arranged in the order of chromosomes, gene density, repeat density and GC content from outside to inside across the 30 pseudochromosomes.

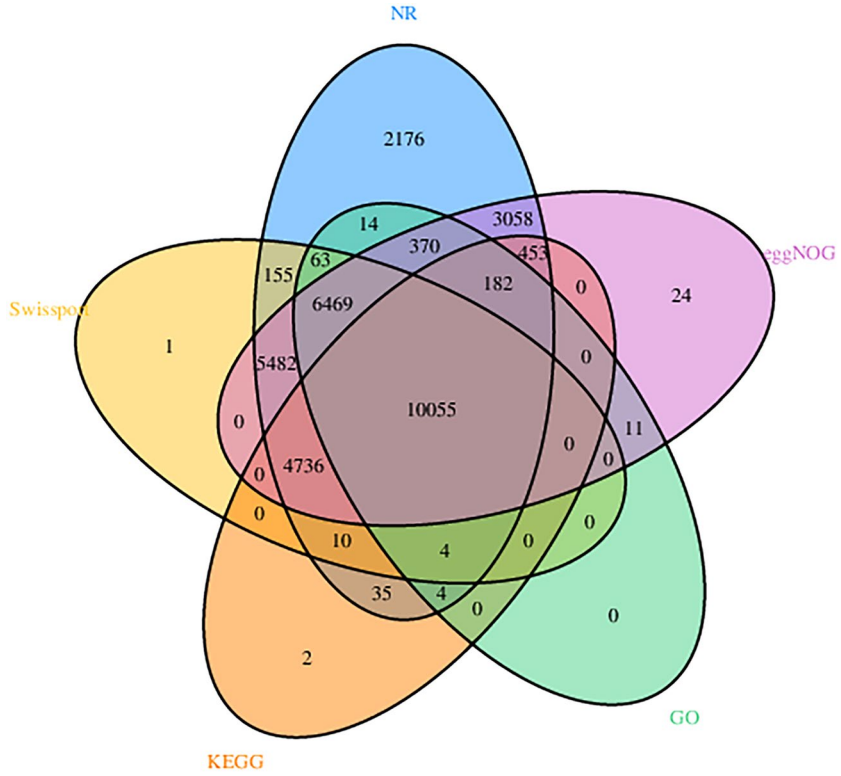

Fig. 6 Venn diagram of gene functional annotations across multiple databases.

LTR retrotransposons dominated the repeat landscape (12.48%), with Gypsy (5.13%) and Copia (1.15%) as the predominant subtypes (Table 2). Non-coding RNAs were annotated using tRNAscan-SE v2.0[39] for tRNAs and INFERNAL v1.1.2[40] against the Rfam v14.1 database[41] for other classes, yielding 1,261 rRNAs, 230 miRNAs, 65 snRNAs, and 467 tRNAs (Supplementary Table S4). Protein-coding genes prediction integrated multiple approaches: homology-based prediction with GeMoMa v1.9[42] using related proteomes and transcript evidence (HISAT2[43] alignments assembled by Cufflinks), and *ab initio* prediction using AUGUSTUS v3.5[23], SNAP[44], GlimmerHMM v3.0.4[45], and GeneMark-ET[46]. Results were integrated using EvidenceModeler, and UTR annotation and alternative splicing analysis were performed using PASA v2.5.2[47]. This process identified 34,541 protein-coding genes with a mean length of 4,949 bp (Table 3). Functional annotation was performed by conducting BLASTP[15] searches against the NR v202108 database[48], Swiss-Prot release 2021_08[49], eggNOG v5.0[50], Gene Ontology (GO)[51], and KEGG PATHWAY[52] databases. A total of 33,304 genes (96.42%) were functionally annotated (Table 3). Subsequent Venn analysis of annotations across these databases identified 10,055 genes with consensus support (Fig. 6).

Data Records

The raw sequencing data generated in this study, including genome survey short reads, PacBio HiFi reads, Hi-C reads, RNA-seq short reads, and Iso-Seq reads, have been deposited in the Genome Sequence Archive (GSA) at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB), under BioProject accession number PRJCA042646[53]. The corresponding dataset accessions are CRA027553[54] (genome survey), CRA027595[55] (PacBio HiFi), CRA027574[56] (Hi-C), CRA027597[57] (RNA-seq), and CRA027598[58] (Iso-Seq). The assembled whole-genome sequence has been deposited in the Genome Warehouse (GWH) at NGDC under accession number GWHGEOG00000000.1[59] and is publicly accessible at https://ngdc.cncb.ac.cn/gwh. The raw sequencing data and genome assembly have also been deposited in the European Nucleotide Archive (ENA) under project accession number PRJEB102980[60], with the genome assembly available under accession number GCA_977063535[61]. The respective ENA accessions for the raw data are: ERR15860102[62] (survey reads); ERR15862836[63] (Iso-Seq reads); ERR15874483[64] (HiFi reads); ERR15873639[65] (Hi-C reads); and ERR15873620[66], ERR15873621[67], ERR15873622[68] and ERR15873623[69] (RNA-seq reads).

Technical Validation

The chromosome-level genome assembly of *Typha minima* (324.66 Mb) was comprehensively validated using orthogonal methods, confirming its high contiguity (Table 1). Hi-C scaffolding anchored 99.65% of the assembly into 30 chromosomes, with chromatin interaction heatmaps demonstrating clear chromosomal compartmentalization (Fig. 4). Assembly completeness was supported by a BUSCO score of 99.2% and K-mer analysis ($k = 23$), which estimated a genome size of 331.55 Mb (Fig. 1)—a deviation of only 2.1% from the final assembly size. High data fidelity was further demonstrated by mapping rates exceeding 98% for both short reads and PacBio

HiFi reads (aligned using BWA v0.7.17), as well as an RNA-seq alignment rate of 98.09%. Rigorous contamination screening—including pre-assembly BLASTN alignment against the NT database (Supplementary Table S3) and post-assembly GC-depth distribution analysis, which showed no aberrant clusters (Fig. 3) —confirmed the absence of detectable foreign sequences. Gene annotation yielded 34,542 protein-coding genes, 96.42% of which were functionally annotated. Among these, 10,055 genes were consistently supported across all databases (Table 3).

Data availability

Raw sequence data have been deposited in the Genome Sequence Archive (GSA) at the National Genomics Data Center (NGDC) under BioProject accession number PRJCA042646[53]. The specific accessions for the genome survey, PacBio HiFi, Hi-C, RNA-seq, and Iso-Seq reads are CRA027553[54], CRA027595[55], CRA027574[56], CRA027597[57], and CRA027598[58], respectively. The genome assembly has been deposited in the Genome Warehouse (GWH) under accession number GWHGEOG00000000.1[59]. All raw data and the assembly are also available in the European Nucleotide Archive (ENA) under project PRJEB102980[60], with the following accessions: ERR15860102[62] (survey); ERR15862836[63] (Iso-Seq); ERR15874483[64] (HiFi); ERR15873639[65] (Hi-C); ERR15873620[66], ERR15873621[67], ERR15873622[68] and ERR15873623[69] (RNA-seq reads).; and GCA_977063535[61] (genome assembly).

Code availability

All bioinformatics tools and software used for genome assembly, annotation, and data analysis in this study were operated strictly according to their official user manuals, with no custom code employed. Software versions and parameters are comprehensively documented in the Methods section.

Received: 15 July 2025; Accepted: 29 December 2025;
Published online: 10 January 2026

References

1. Bansal, S. *et al.* Typha (Cattail) invasion in North American wetlands: biology, regional problems, impacts, ecosystem services, and management. *Wetlands* **39**, 645–684 (2019).
2. Carpenter, S. R. & Lodge, D. M. Effects of submersed macrophytes on ecosystem processes. *Aquatic Botany* **26**, 341–370 (1986).
3. Lewis, M. & Thursby, G. Aquatic plants: Test species sensitivity and minimum data requirement evaluations for chemical risk assessments and aquatic life criteria development for the USA. *Environ. Pollut.* **238**, 270–280 (2018).
4. Thomaz, S. M. & Cunha, E. R. The role of macrophytes in habitat structuring in aquatic ecosystems: methods of measurement, causes and consequences on animal assemblages' composition and biodiversity. *Acta Limnol. Bras.* **22**, 218–236 (2010).
5. Alufasi, R. *et al.* Internalisation of Salmonella spp. by Typha latifolia and Cyperus papyrus *in vitro* and implications for pathogen removal in Constructed Wetlands. *Environ. Technol.* **43**, 949–961 (2022).
6. National Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China (2020 edition): Volume I. China Medical Science Press, Beijing (2020).
7. Smith, S. G. Typha: its taxonomy and the ecological significance of hybrids. *Arch. Hydrobiol* **27**, 129–138 (1987).
8. Csencsics, D. *et al.* La petite massette: Habitant menacé d'un biotope rare. Notice pour le praticien 43. Institut fédéral de recherches WSL, Birmensdorf (2008).
9. Zhou, B. *et al.* Revised phylogeny and historical biogeography of the cosmopolitan aquatic plant genus Typha (Typhaceae). *Sci. Rep.* **8**, 8813 (2018).
10. Liao, Y. *et al.* Chromosome-level genome and high nitrogen stress response of the widespread and ecologically important wetland plant *Typha angustifolia*. *Front. Plant Sci.* **14**, 1138498 (2023).
11. Widanagama, S. D., Freeland, J. R., Xu, X. & Shafer, A. B. Genome assembly, annotation, and comparative analysis of the cattail *Typha latifolia*. *G3: Genes, Genomes, Genetics* **12**(2), jkab401 (2022).
12. Aleman, A. *et al.* Development of genomic resources for cattails (*Typha*), a globally important macrophyte genus. *Freshwater Biology* **69**(1), 74–83 (2024).
13. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
14. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
15. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
17. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
18. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications.* **11**(1), 1432 (2020).
19. 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 (2021).
20. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
21. 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 (2021).
22. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
23. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Combining gene prediction methods with alignment information in the AUGUSTUS gene finder. *Bioinformatics* **22**, 417–425 (2006).
24. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* **25**(14), 1754–1760 (2009).
26. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
27. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
28. Marc van Dijk, M. & Bonvin, A. M. 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.* **37**, W235–W239 (2009).

29. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258.e1 (2018).
30. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).
31. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
32. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
33. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
34. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
35. 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).
36. Tempel, S., Jurka, M. & Jurka, J. VisualRepbase: an interface for the study of occurrences of transposable element families. *BMC Bioinform.* **9**, 345 (2008).
37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **25**, 4.10.1–4.10.14 (2009).
38. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
39. 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).
40. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
41. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
42. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods Mol. Biol.* **1962**, 161–177 (2019).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
44. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
45. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
46. Bruna, T., Lomsadze, A. & Borodovsky, M. A new gene finding tool GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. bioRxiv 2023.01.13.524024 (2024).
47. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
48. Sayers, E. W. *et al*. GenBank. *Nucleic Acids Res.* **49**, D92–D96 (2021).
49. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
50. 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).
51. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
52. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
53. *NGDC BioProject* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA042646 (2024).
54. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA027553 (2024).
55. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA027595 (2024).
56. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA027574 (2024).
57. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA027597 (2024).
58. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA027598 (2024).
59. *NGDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/98240/show (2024).
60. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB102980 (2025).
61. *European Nucleotide Archive* https://identifiers.org/insdc.gca:GCA_977063535 (2025).
62. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15860102 (2025).
63. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15862836 (2025).
64. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15874483 (2025).
65. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15873639 (2025).
66. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15873620 (2025).
67. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15873621 (2025).
68. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15873622 (2025).
69. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERR15873623 (2025).

Acknowledgements

This study was supported by the National Water Pollution Control and Treatment Science and Technology Major Project, China (No. 2015ZX07503005). The calculations in this paper were performed using the supercomputing system at the Supercomputing Center of Wuhan University.

Author contributions

X.X. designed the research; L.H. carried out the field collections; J.D. carried out the experiments and performed the data analysis; J.D., L.H. and X.X. wrote and revised the manuscript. All authors read and approved the manuscript.

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/s41597-026-06547-2.

Correspondence and requests for materials should be addressed to X.X.