scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of watershield (*Brasenia schreberi*)

Bei Lu[1,2], Tao Shi[1] & Jinming Chen[1]✉

Watershield (*Brasenia schreberi*) is an aquatic plant that belongs to the basal angiosperm family Cabombaceae. This species has been cultivated as an aquatic vegetable for more than 3000 years in East Asia, but the natural populations have greatly declined in recent decades and have become endangered in several countries of East Asia. In this study, by using PacBio long reads, Illumina short reads, and Hi-C sequencing data, we assembled the genome of *B. schreberi*, which was approximately 1170.4 Mb in size with a contig N50 of 7.1 Mb. Of the total assembled sequences, 93.6% were anchored to 36 pseudochromosomes with a scaffold N50 of 28.9 Mb. A total of 74,699 protein-coding genes were predicted in the *B. schreberi* genome, and 558 Mb of repetitive elements occupying 47.69% of the genome were identified. BUSCO analysis yielded a completeness score of 95.8%. The assembled high-quality genome of *B. schreberi* will be a valuable reference for the study of conservation, evolution and molecular breeding in this species.

Background & Summary

Brasenia schreberi J.F. Gmel. (watershield) is a monotypic species in the genus *Brasenia* (Cabombaeae), which belongs to the basal angiosperm order Nymphaeales. This species is a perennial floating leaved freshwater aquatic plant that is found in the tropical and temperate regions of America, Africa, Australasia, and Asia[1]. *B. schreberi* produces thick mucilage that covers the juvenile leaf abaxial surface and buds[2,3] (Fig. 1). This mucilage has been found to have anti-algal and antibacterial properties and may function as an herbivory defense, protecting young buds from abrasion, and as an excellent biological lubricant[4]. *B. schreberi* has been cultivated as an aquatic vegetable for more than 3000 years in East Asia due to the importance of mucilage-covered young leaves and buds in the diet and the special flavor of the mucilage[5].

Plant mucilage is a gelatinous matrix comprising mostly polysaccharides known as pectins produced by glandular trichomes (GTs), seed coats, root hairs, etc.[6,7], serving various functions for plants. Although all investigated lineages of the basal angiosperm order Nymphaeales possess epidermal trichome-like structures (GTs), only a few species, such as *B. schreberi*, have a mucilage layer secreted by the GTs[7–9]. Thus, *B. schreberi* represents an interesting system for studying the evolution and molecular mechanisms of plant mucilages. In addition, plant GTs have been important target traits for crop breeding[10].

In the past three decades, due to human activity and habitat loss, the natural populations of *B. schreberi* have decreased significantly and are considered endangered in several counties of East Asia[11,12]. For example, in China, this species has been previously listed as the first category of key protected wild plants[13]; in Korea, it is listed as a critically endangered species[14]. For conservation purposes, several population genetic studies using few molecular markers have been conducted on *B. schreberi* in China and Korea[11,12,15,16]. However, these studies utilized only limited regions of the genome, and further conservation genomics studies at the whole genome scale are needed to establish effective management strategies for this endangered species in Asia.

In this study, we presented a high-quality genome sequence for *B. schreberi* obtained using PacBio, Illumina, and Hi-C technologies (Fig. 1). The assembled genome had a size of 1,170.4 Mb with a contig N50 of 7.1 Mb and a scaffold N50 of 28.9 Mb (Table 1). The assembled scaffolds were further anchored to 36 pseudochromosomes, with an anchoring rate of 93.6% (Fig. 2, Table 1). A total of 74,699 protein-coding gene models were fully annotated (Table 2). Repetitive elements (TEs), with a collective length of 558 Mb, occupied 47.69% of the *B. schreberi* genome (Table 2). The quality of the final genomic assembly was assessed to be high gene completeness (95.8%),

[1]Aquatic Plant Research Center, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, 4300074, China. [2]University of Chinese Academy of Sciences, Beijing, 100049, China. ✉e-mail: jmchen@wbgcas.cn

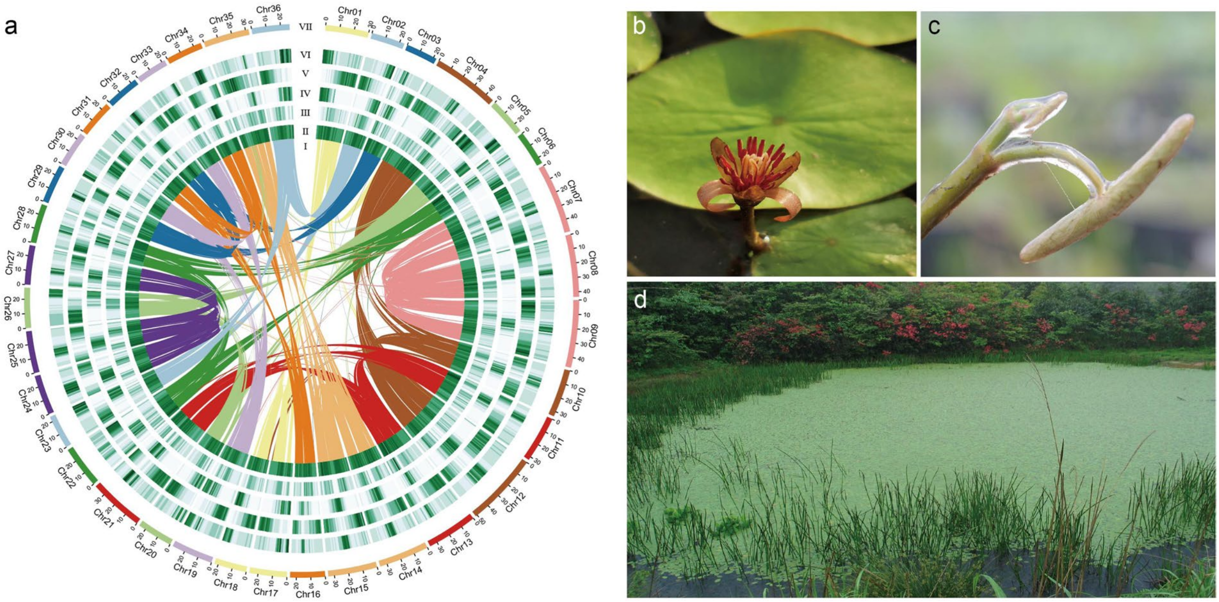

Fig. 1 Overview of the genome and plants of *Brasenia schreberi*. (**a**) Circos plot of anchored *B. schreberi* genomic features. Difference tracks show: (I) Links of intragenomic syntenic blocks, only blocks with more than five syntenic genes are shown, (II) LTR/Copia density in 1 Mb sliding windows, (III) LTR/Gypsy density in 1 Mb sliding windows, (IV) repeat density, (V) gene density in 1 Mb sliding windows (minimum-maximum, 0–30), (VI) GC content in 1 Mb sliding windows. (**b**) A flower on its second day of blossoming. (**c**) Juvenile leaf and bud with mucilage covering. (**d**) A natural population during the growing season.

Genome assembly statistics	
Total length of Contigs (Mb)	1,170.4
Number of Contigs	3,846
N50 length of contigs (Mb)	7.1
Longest contig (Mb)	51.2
Number of scaffolds	2,173
N50 length of scaffolds (Mb)	28.9
Anchored rate (%)	93.6
BUSCO score (Eukaryota) (%)	95.8
LTR Assembly Index, LAI	6.21

Table 1. Summary of *Brasenia scherberi* genome assembly.

as indicated by BUSCO[17]. The assembled high-quality genome of *B. schreberi* should be a valuable resource for future conservation genomics studies. In addition, our assembled reference genome of this basal angiosperm offers a new resource for studying the origin and early adaptive evolution of angiosperms and for revealing the molecular basis of the trait of mucilage secretion, which will facilitate molecular breeding in this aquatic vegetable.

Methods

Sampling, sequencing and genome size estimation. *B. schreberi* plants were originally collected from a natural population in Lichuan, Hubei Province, China, in 2018 and cultivated in the Wuhan Botanical Garden (WBG) of the Chinese Academy of Sciences. After the collection of juvenile leaves, the mucilage on the back of the leaves was washed off, and the leave samples were promptly stored in liquid nitrogen. Then the high-quality genomic DNA was extracted from the processed samples using the MagicMag plant genomic DNA Micro Kit (Sangon Biotech Co.) and used for subsequent Illumina and PacBio sequencing. The MGI libraries were constructed and sequenced on a DNBSEQ-T7 platform at an expected coverage of 80 × (see table deposited at Figshare[18]). The MGI short reads were used for both genome size estimation and residual error correction in the *de novo* genome assembly. For PacBio sequencing, 20 kb DNA libraries were constructed and then sequenced using single molecule real-time (SMRT). A total of 101 Gb of data composed of 5.4 million subreads were generated on the PacBio Sequel platform (Pacific Biosciences) (see table deposited at Figshare[18]). The Hi-C libraries were constructed following a previously published protocol[19]. The sample underwent liquid nitrogen grinding and was then cross-linked with 4% formaldehyde at room temperature under vacuum for 30 minutes. Quenching

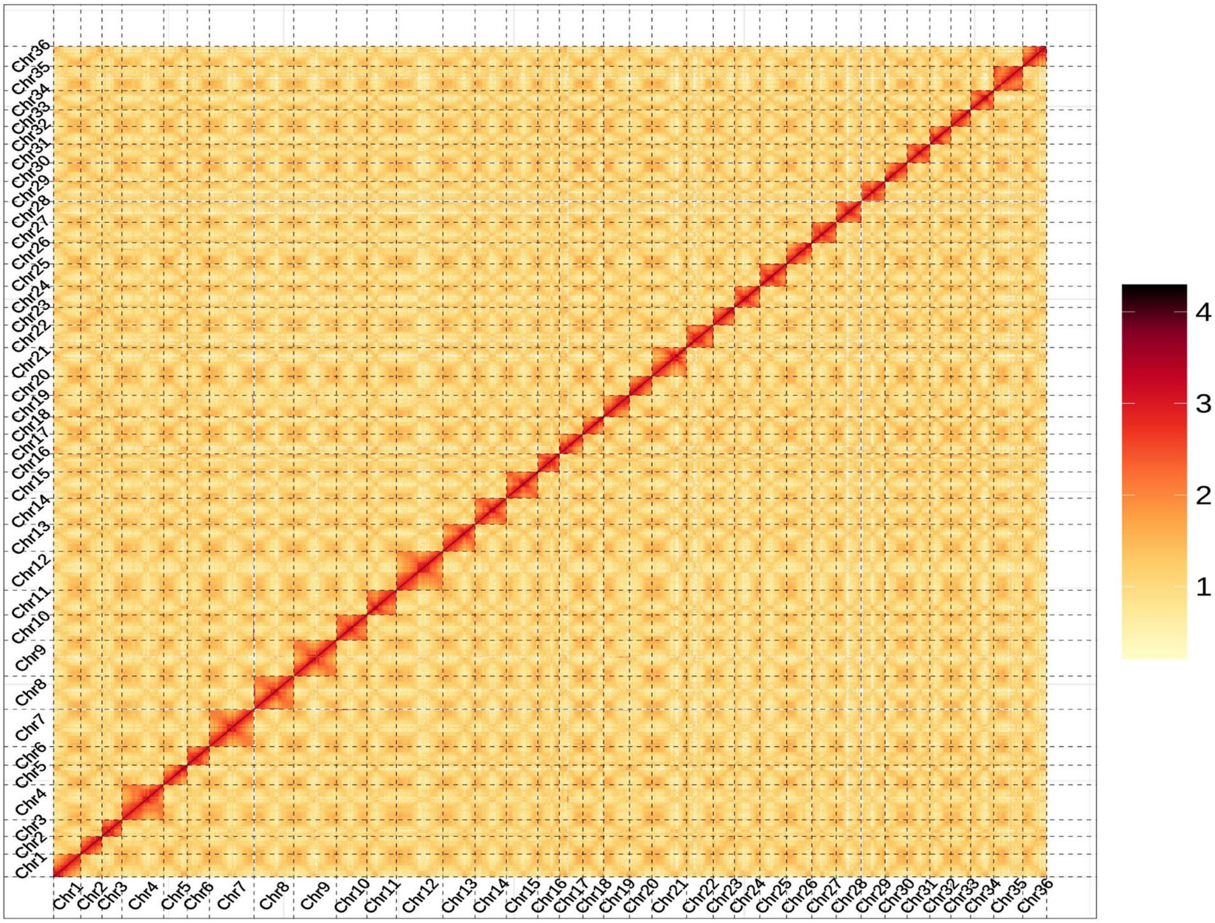

Fig. 2 Hi-C interaction heatmap within pseudochromosomes of *Brasenia schreberi*.

Genome assembly statistics	Count	Proportion in genome (%)
Predicted protein-coding genes	74,699	
NR	61,366	82.15
pfam	47,317	63.34
KEGG	29,431	39.40
GO	41,427	55.46
eggNOG	55,144	73.82
Annotated	61,618	82.49
Percentage of repeat elements	1,123,230	47.69
Copia	165,769	11.23
Gypsy	156,404	14.18

Table 2. Summary of *Brasenia scherberi* genome annotations.

of the crosslinking reaction was achieved by adding 2.5 M glycine for 5 minutes followed by incubation on ice for 15 minutes. After centrifugation at 2500 rpm and 4 °C for 10 minutes, the pellet was washed with 500 μl PBS, centrifuged again, and resuspended in a lysis buffer. The resulting supernatant was subjected to further centrifugation, and the pellet was washed, resuspended, and solubilized using dilute SDS at 65 °C for 10 minutes. Subsequent steps involved digestion with a 4-cutter restriction enzyme DpnII overnight at 37 °C, marking of DNA ends with biotin-14-dCTP, and blunt-end ligation of cross-linked fragments. Proximal chromatin DNA was re-ligated, nuclear complexes were reversely cross-linked with proteinase K at 65 °C, and DNA was purified by phenol-chloroform extraction. Biotin was removed from nonligated fragment ends, and sheared fragments were repaired. Biotin-labeled Hi-C samples were enriched using streptavidin C1 magnetic beads. After addition of A-tails and ligation with Illumina PE sequencing adapters, Hi-C libraries were PCR-amplified (12–14 cycles) and sequenced on Illumina PE150 platform at Novogene Biotech Co., Ltd. (Beijing, China) for chromosome construction.

For genome functional annotation, transcriptome sequencing was performed with seven tissues of *B. schreberi*, including stamen, pistil, perianth, stem, root, rhizome, and leaf. RNA libraries were prepared using the TruSeq RNA Sample Prep Kit (Illumina, USA) according to the manufacturer's instructions, and PE150 sequencing was conducted on an Illumina NovaSeq. 6000 platform at Novogene Biotech Co., Ltd. (Beijing, China).

The genome size and ploidy levels of *B. schreberi* were estimated using two methods: (i) flow cytometry, which was conducted on a BD AccuriTMC6 flow cytometer (BD Biosciences) using the leaf of *Nelumbo nucifera* (genome size ≈ 807.6 Mb[20]) as a reference, and the genome size of *B. schreberi* was estimated as ~1100 megabases (Mb) by this method (see figure deposited at Figshare[18]); and (ii) *k*-mer-based estimation, in which the *k*-mer distribution of Illumina reads was counted by using jellyfish v2.3.0 (*k*-mer = 21, parameters: count -m 21 -t 10 -s 1 G), and then the genome size and the rate of heterozygosity were estimated to be ~956.2 Mb and 0.10%, respectively, by GenomeScope online version (http://qb.cshl.edu/genomescope/) using the *k*-mer count distribution file (see figure deposited at Figshare[18]); ploidy levels were assessed by Smudgeplot v0.2.5[21] based on heterozygous *k*-mer pairs (see figure deposited at Figshare[18]).

Genome assembly. Canu v.1.8[22] (parameters: -p out genomeSize = 1.5 g maxThreads = 30 useGrid = false) was used for self-correction, trimming, and assembly. To polish the draft assembly, PacBio subreads were subjected to three rounds of polishing with the program Racon v1.4.3 (https://github.com/isovic/racon), and then the Illumina paired-end reads were further subjected to three rounds of polishing with the program Pilon v1.23[23] (parameters: --fix all --changes). Finally, the total length of the draft genome was 1170.4 Mb, comprising 3,846 contigs with a contig N50 of 7.1 Mb and a maximum length of 51.2 Mb (Table 1).

The assembly was refined using high-throughput chromosome conformation capture (Hi-C) data. The 2,802,461 Hi-C paired-end reads, which were grouped to the chromosome level using ALLHiC[24], were remapped to the draft assembly. We divided the assembled chromosomes into equally sized bins (500 Kb) and constructed an interaction heatmap based on the number of valid paired-end reads supporting interactions between each pair of bins. Then, the visual correction of assembly was finalized using JuiceBox v.2.1.10[25] based on the intensity of chromosome interaction (Fig. 2). The specific criteria we employed for visual correction were as follows: We adjusted the assembly based on the principle that intra-chromosomal interactions should be stronger than inter-chromosomal interactions. If there were evident assembly errors within a completed contig, we would break and adjust it according to the interaction relationships. Additionally, very short contigs without any interaction relationships were placed in the unassigned category. The chromosome-level genome assembly was improved, containing 2,173 scaffolds with a scaffold N50 of 28.9 Mb (Table 1). These scaffolds were further anchored onto 36 pseudochromosomes[26–28], resulting in a total of 36 chromosomes and 2137 additional scaffolds, with an anchor rate of 93.6% (Table 1).

Genome annotation. First, the EDTA (Extensive *de novo* TE Annotator) program v2.0.1[29] was used to annotate the repeat sequences, including the repetitive element (TE) sequences, and generate the masked repeat sequence for gene prediction. Repetitive elements with a collective length of 558 Mb occupied 47.69% of the *B. schreberi* genome (Table 2). Then, three algorithms were used to predict genes: *ab initio*, homolog, and transcriptome alignment. Seven RNAseqs generated in this study were used in both methods of transcriptome-based alignment: (1) Transcriptomes were assembled by Trinity v2.5.1[30] (both *de novo* and guided), and the PASA (Program to Assemble Spliced Alignments v2.3.3)[31] program was used to align these redundant transcriptomes to genome sequences; (2) transcriptomes were assembled from a hisat2-stringtie pipeline[32], and the open reading frames (ORFs) were predicted by TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/). Homolog protein comparison was conducted using GETA v2.4.12 (https://github.com/chenlianfu/geta) (parameters: homolog_genewise–cpu 40–coverage_ratio 0.4–evalue 1e-9–max_gene_length 2000) with the program GeneWise[33]. The *ab initio* method was Braker2[34]. The EVidenceModeler program (EVM-1.1.1)[35] was used to integrate the above redundant annotation information. Three rounds of PASA annotation updates were performed to obtain annotation information for the genome. Combining the *ab initio*, RNA-seq, and homology-based methods, a total of 74699 protein-coding gene models were fully annotated (Table 2). The predicted gene length overlap larger than 30% with repeat sequences was filtered by TransposonPSI v1.0.0 (http://transposonpsi.sourceforge.net/) for the downstream analysis. For functional annotation, we performed searches of our predicted protein-coding genes against the non-redundant (NR) using BLASTP v2.9.033[36], Pfam, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, as well as the eggNOG database using EggNOG-mapper[37] v2.1.11 (Table 2). The completeness 82.15% of protein-coding genes had significant hits in the functional annotation databases (Table 2).

Orthologue and phylogenetic analyses. The genome protein sequences of 13 angiosperms were used to determine the orthologs using OrthoFinder v2.5.4[38], including three Nymphaeales (*B. schreberi*, *Nymphaea colorata*, and *Euryale ferox*), three monocots (*Acorus tatarinowii*, *Zostera marina*, and *Oryza sativa*), three magnoliids (*Aristolochia fimbriata*, *Cinnamomum kanehirae*, and *Magnolia biondii*), three eudicots (*Aquilegia coerulea*, *Vitis vinifera*, and *Arabidopsis thaliana*), and *Amborella trichopoda* (see table deposited at Figshare[18]). The resulted 158 single-copy orthologs were aligned using MAFFT v7.505 with default settings[39]. The corresponding nucleotide sequence alignments of the protein alignments were extracted using pal2nal.pl v14[40] and trimmed with Gblocks v0.91b[41] with the codon model. The maximum likelihood tree was constructed under 'GTRGAMMA' model of nucleotide substitution using RAxML v8.2.12[42] with 100 bootstrap replicates (Fig. 3).

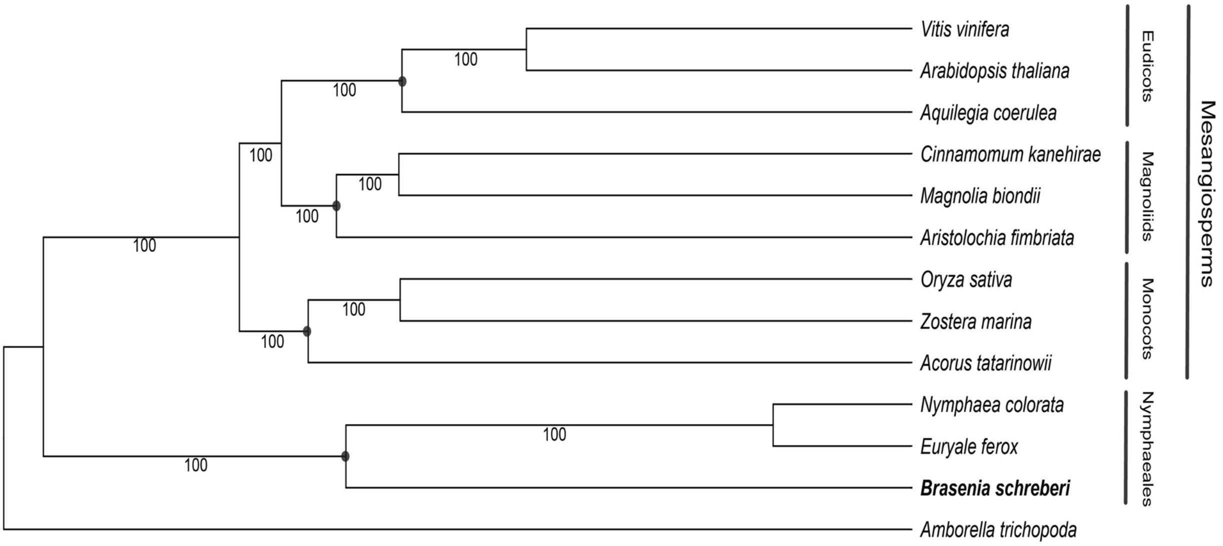

Fig. 3 Phylogenetic tree of 13 angiosperm species generated by RAxML, number below branches displaying bootstraps support of 100%.

Data Records

The raw data of MGI and Hi-C sequencing were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive database with accession numbers SRR24223717[43], and SRR24223715[44]. Seven transcriptome data were submitted to NCBI with accession numbers SRR24136212[45], SRR24136211[46], SRR24136210[47], SRR24136209[48], SRR24136208[49], SRR24136207[50], SRR24136206[51], under the BioProject accession number PRJNA954463. The genome assembly data, genome annotation files, gene CDS, and protein data have been deposited into CNGB Sequence Archive (CNSA)[52] with the accession number CNA0069000 under the BioProject accession number CNP0004217. The raw PacBio sequences have been deposited into CNSA[52] under the BioProject accession number CNP0004217 (https://ftp.cngb.org/pub/CNSA/data5/CNP0004217/CNS0724876/CNX0616196/CNR0710381/). The genome annotation files had also been deposited at the Figshare[53]. The genome genome assembly data had also been submitted to GenBank with accession number JARYZE000000000[54].

Technical Validation

Three methods were used to evaluate the quality of the genome assembly. First, in the Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.2.2)[17] evaluation, complete and single-copy, complete and duplicated, fragmented, and missing categories accounted for 44.1% (712), 51.7% (835), 1.9% (30), and 2.3% (37) of 1614 Eukaryota BUSCO genes identified in the chromosome-level genome assembly, respectively. Then, we calculated the long terminal repeat (LTR) assembly index (LAI) based on the EDTA results with the default settings. The LTR Assembly Index (LAI) is used as a validation measure to assess the quality of LTR sequences in genome assembly. LAI calculates the number of correctly positioned LTRs and considers the integrity of LTR sequences. Higher LAI values indicate better positioning and integrity, implying higher assembly quality and accuracy of LTR sequences. It complements other assembly quality metrics, providing a comprehensive evaluation of the assembly outcomes. We also calculated the mapping rate of seven transcriptomes generated from different tissues and developmental stages (see table deposited at Figshare[18]). In total, the quality of the final genomic assembly was assessed to be high completeness (95.8% indicated by BUSCO), contiguity (6.21 indicated by LAI), and consistency (97%~98% mapping rate of RNA-seq datasets).

Code availability

No custom programming or coding was used.

Received: 27 April 2023; Accepted: 11 July 2023;
Published online: 19 July 2023

References

1. Osborn, J. M. & Schneider, E. L. Morphological studies of the Nymphaeaceae Sensu Lato. XVI. The floral biology of *Brasenia schreberi*. *Ann. Mo. Bot. Gard.* **75**, 778, https://doi.org/10.2307/2399366 (1988).
2. Elakovich, S. D. & Wooten, J. W. An examination of the phytotoxicity of the water shield, *Brasenia schreberi*. *J. Chem. Ecol.* **13**, 1935–1940, https://doi.org/10.1007/BF01014676 (1987).
3. Yang, C. D., Zhang, X., Zhang, F., Wang, X. E. & Wang, Q. F. Structure and ion physiology of *Brasenia schreberi* glandular trichomes *in vivo*. *PeerJ* **7**, e7288, https://doi.org/10.7717/peerj.7288 (2019).
4. Thompson, K. A., Sora, D. M., Cross, K. S., St. Germain, J. M. & Cottenie, K. Mucilage reduces leaf herbivory in Schreber's watershield, *Brasenia schreberi* J.F. Gmel. (Cabombaceae). *Botany* **92**, 412–416, https://doi.org/10.1139/cjb-2013-0296 (2014).
5. Xie, C. *et al*. Environmental factors influencing mucilage accumulation of the endangered *Brasenia schreberi* in China. *Sci. Rep.* **8**. https://doi.org/10.1038/s41598-018-36448-3 (2018)

6. Fahn, A. Secretory tissues in vascular plants. *New Phytol.* **108**, 229–257, https://doi.org/10.1111/j.1469-8137.1988.tb04159.x (1988).
7. Kordyum, E., Mosyakin, S., Ivanenko, G., Ovcharenko, Y. & Brykov, V. Hydropotes of young and mature leaves in *Nuphar lutea* and *Nymphaea alba* (Nymphaeaceae): Formation, functions and phylogeny. *Aquat. Bot.* **169**, https://doi.org/10.1016/j.aquabot.2020.103342 (2021).
8. Carpenter, K. J. Specialized structures in the leaf epidermis of basal angiosperms: morphology, distribution, and homology. *Am. J. Bot.* **93**, 665–681, https://doi.org/10.3732/ajb.93.5.665 (2006).
9. Tozin, L. R. D. S. & Rodrigues, T. M. Revisiting hydropotes of Nymphaeaceae: ultrastructural features associated with glandular functions. *Acta Botanica Brasilica* **34**, 31–39, https://doi.org/10.1590/0102-33062019abb0120 (2020).
10. Glas, J. *et al.* Plant Glandular Trichomes as targets for breeding or engineering of resistance to herbivores. *Int. J. Mol. Sci.* **13**, 17077–17103, https://doi.org/10.3390/ijms131217077 (2012).
11. Kim, C. *et al.* Population genetic structure of the endangered *Brasenia schreberi* in South Korea based on nuclear ribosomal spacer and chloroplast DNA sequences. *J. Plant Biol.* **55**, 81–91, https://doi.org/10.1007/s12374-011-9193-4 (2012).
12. Li, Z. Z., Gichira, A. W., Wang, Q. F. & Chen, J. M. Genetic diversity and population structure of the endangered basal angiosperm *Brasenia schreberi* (Cabombaceae) in China. *PeerJ* **6**, e5296, https://doi.org/10.7717/peerj.5296 (2018).
13. Yu, Y. F. A milestone of wild plant conservation in China. *Plants* **5**, 3–11 (1999).
14. Lee, H.W. *et al.* Categorization and conservation of the threatened plant species in environmental impact assessment. *Korea Environment Institute, Seoul* (2005).
15. Zhang, G. F. & Gao, B. Q. Analysis on genetic diversity and genetic structure of *Brasenia schreberi* in Jiangsu and Zhejiang Provinces revealed by ISSR markers. *J. Lake Sci.* **20**, 662–668 (2008).
16. Kim, C. K., Na, H. R. & Choi, H. K. Conservation genetics of endangered *Brasenia schreberi* based on RAPD and AFLP markers. *J. Plant Biol.* **51**, 260–268, https://doi.org/10.1007/BF03036125 (2008).
17. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
18. Lu, B. Supplementary figures and tables. *figshare.* https://doi.org/10.6084/m9.figshare.22567210 (2023).
19. Belton, J. M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
20. Shi, T. *et al.* Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plant. *Mol. Biol. Evol.* **37**, 2394–2413, https://doi.org/10.1093/molbev/msaa105 (2020).
21. 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).
22. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736, https://doi.org/10.1101/gr.215087.116 (2017).
23. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
24. Zhang, X. T., Zhang, S. C., Zhao, Q., Ming, R. & Tang, H. B. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
25. Robinson, J. T. *et al.* Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Systems* **6**, 256–258. e251, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
26. Wei, P. H., Chen, W. P. & Chen, R. Y. Study on the karyotype analysis of Nymphaeaceae and its taxonomic position. *J. Syst. Evol.* **32**, 293–300 (1994).
27. Chen, R. *et al.* Chromosome atlas of various bamboo species. Chromosome atlas of major economic plants genome in China II (Beijing: Science Press, 2002).
28. Diao, Y. *et al.* Nuclear DNA C-values in 12 species in Nymphaeales. *Caryologia* **59**, 25–30, https://doi.org/10.1080/00087114.2006.10797894 (2006).
29. Ou, S. J. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
30. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
31. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
32. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667, https://doi.org/10.1038/nprot.2016.095 (2016).
33. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
34. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, https://doi.org/10.1093/nargab/lqaa108 (2021).
35. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
37. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
38. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238, https://doi.org/10.1186/s13059-019-1832-y (2019).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780, https://doi.org/10.1093/molbev/mst010 (2013).
40. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–612, https://doi.org/10.1093/nar/gkl315 (2006).
41. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552, https://doi.org/10.1093/oxfordjournals.molbev.a026334 (2000).
42. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24223717 (2023).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24223715 (2023).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136212 (2023).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136211 (2023).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136210 (2023).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136209 (2023).

49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136208 (2023).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136207 (2023).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24136206 (2023).
52. Guo, X. Q. *et al*. CNSA: a data repository for archiving omics data. *Database (Oxford)* **2020**, 1–6, https://doi.org/10.1093/database/baaa055 (2020).
53. Lu, B. Annotation of *Brasenia* genome. *figshare*. https://doi.org/10.6084/m9.figshare.22591369.v1 (2023).
54. Lu, B. The genome information of Brasenia. *GenBank* https://identifiers.org/ncbi/insdc:JARYZE000000000 (2023).

Acknowledgements

This study was financially supported by the Biological Resources Program, CAS (No. KFJ-BRP-007-009) and the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2019335).

Author contributions

J.C. conceived the idea, supervised the work, and revised the manuscript. B.L. and J.C. prepared the plant materials. B.L. and T.S. analyzed the data. B.L. and J.C. wrote the original draft and revised the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.