Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Draft genome of Brasenia schreberi, a worldwide distributed and endangered aquatic plant

Lin-Fang Wu
Guangzhou Linfang Ecological Technology Co., Ltd
Wei-Guang Zhu
Chinese Academy of Sciences
En-Ping Yu
Chinese Academy of Sciences
Hong-Lin Cao
Chinese Academy of Sciences
Zheng-Feng Wang
wzf@scib.ac.cn

Chinese Academy of Sciences

Data Note

Keywords: gene prediction, genome feature, genome annotation, repetitive sequences

Posted Date: February 2nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3915726/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomic Data on March 4th, 2024.
See the published version at https://doi.org/10.1186/s12863-024-01212-2.

Abstract

Objectives

Brasenia is a monotypic genus in the family of Cabombaceae. The only species, *B. schreberi*, is a macrophyte distributed worldwide. Because it requires good water quality, it is endangered in China and other countries due to the deterioration of aquatic habitats. The young leaves and stems of *B. schreberi* are covered by thick mucilage, which has high medical value. As an allelopathic aquatic plant, it can also be used in the management of aquatic weeds. Here, we present its assembled and annotated genome to help shed light on medial and allelopathic substrates and facilitate their conservation.

Data description:

Genomic DNA and RNA extracted from *B. schreberi* leaf tissues were used for whole genome and RNA sequencing using a Nanopore and/or MGI sequencer. The assembly was 1,055,148,839 bp in length, with 92 contigs and an N50 of 22,379,495 bp. The repetitive elements in the assembly were 555,442,205 bp. A completeness assessment of the assembly with BUSCO and compleasm indicated 88.4 and 90.9% completeness in the Eudicots database and 95.4 and 96.6% completeness in the Embryphyta database. Gene annotation revealed 67,747 genes that coded for 73,344 proteins.

Objective

Brasenia schreberi is an aquatic and perennial herb in the Cabombaceae family. It is a monotypic species with oval-shaped leaves that can submerge or float on the water's surface, similar to water lilies. Except for Europe and Antarctica, it is currently distributed on all continents of the world [1]. However, palaeobotanical records indicate that *B. schreberi* was a frequent element in Europe before the last glacial period [1]. Its habitats include ponds, lakes, and sluggish streams, but they must be clean and acidic and have nutrient-enriched sediment [1, 2]. Due to the deterioration of water quality and habitat loss, it is listed at the second level of national key protected wild plants in China and is endangered in other countries [2, 3]. Its edible young leaves and stems are coated with a thick mucilage that is mainly composed of polysaccharides and has high medical value [2, 4, 5]. Mucilage has been confirmed to be a defense against herbivores and bacteria [3, 6]. *Brasenia schreberi* contains allelopathic components that can be used in the management of aquatic weeds [7]. As important values in taxonomy, ecology, and economy and in its endangered situation, a genome assembly was published previously [8] for its better conservation and breeding. However, given its wide distribution worldwide and existing substantial genetic diversities [3, 9], we present an alternative *B. schreberi* genome to better understand its evolution and adaptation and to enhance its conservation, management, and utility in the future.

Data description

Leaf samples of *B. schreberi* were collected from an individual planted in the South China Botanical Garden, Guangzhou, China. The DNA or RNA extracted from its leaf tissues was used to construct three sequencing libraries, including long read whole genome sequencing (WGS) using a Nanopore PromethION sequencer, short read WGS using an MGI DNBSEQ-T7 sequencer, and RNA sequencing (RNA-seq) using an MGI DNBSEQ-T7 sequencer. Under the MGI platforms, a 150 bp paired-end mode was applied for both short WGS and RNA-seq. The long-read WGS generated about 113.0 GB of data (Data file 1) [10], short-read WGS generated about 130.6 GB data (Data file 2) [11], and RNA-seq generated about 27.6 GB data (Data file 3) [12].

After sequencing, short WGS reads were trimmed by Sickle v1.33 [13] using the parameter "-q 30 -l 80". KmerGenie v1.7044 [14] (under the parameter of "-k 141 --diploid") was then used to estimate the genome size of *B. schreberi* with trimmed short WGS reads. The estimated genome size was 963,304,542 bp. Porchop v0.2.4 [15] and ontbc v1.1 [16] were used to remove adapter and low-quality reads (scores < 7 and lengths < 5000 bp) in long WGS reads. NextDenovo v2.3.1 [17] was then used to assemble the genome with the filtered long reads. Pseudohaploid [18] and Purge_Dups v1.2.6 [19] were applied to remove redundant contigs. Subsequently, Racon v1.5.0 [20], hapo-G v1.3.2 [21], and polypolish v0.5.0 [22] were used to polish the assembly. The final assembly was 1,055,148,839 bp in length, with 92 contigs and a contig N50 of 22,379,495 bp (Data file 4) [23]. BUSCO v5.5.0 [24] and compleasm v0.2.5 [25] were used to assess the completeness of the assembly with Eudicots odb10-2020-09-10 and Embryphyta odb10 2020-09-10 databases. BUSCO revealed 88.4 and 95.4% completeness in the Eudicots and Embryphyta databases, respectively (Data files 5–6) [26, 27]. Compleasm revealed 90.9 and 96.7% completeness in the Eudicots and Embryphyta databases, respectively (Data files 7–8) [28, 29].

Repetitive elements in the *B. schreberi* assembly were estimated by RED v2.0 [30] and EDTA v2.1.3 [31], which revealed 452,408,938 (Data file 8) [32] and 521,424,853 bp (Data file 9) [33] of sequences, respectively. Combining the RED and EDTA results revealed 555,442,205 bp of repetitive sequences (Data file 10) [34], which were used to soft-mask the assembly. Braker3 v.3.0.6 [35] was used to predict the primary gene structures using transcriptome data and reference protein sequences (Data file 11) [36]. The Braker results were then incorporated into the Funannotate pipeline v1.8.16 [37] to obtain integrated gene sets. The pipeline included four steps: "train", "predict", "update", and "annotate". For the former three steps, the parameter "--max_intronlen 1000000" was used, while in the "predict" step, the parameters "--busco_seed_species arabidopsis --organism other --busco_db embryophyta" were added. The fourth "annotate" step was used for gene function annotation. The final gene prediction obtained 67,747 protein-coding genes and 813 tRNA genes (Data files 12–14) [38–40]. Functional annotation of protein-coding genes is shown in Data files 15–16 [41, 42].

Limitations

The current *B. schreberi* assembly in this study is fragmented. Future sequencing technologies, including Hi-C, Nanopore ultra-long sequencing, PacBio HiFi, 10X Genomics linked sequencing, and Bionano optical maps, are needed for complete and gapless genome assembly.

However, our assembly displayed a completeness comparable to the previously reported *B. schreberi* assembly [8], which showed 89.0 and 95.9% completeness using BUSCO in the Eudicots and Embryphyta databases, respectively, and 91.3% and 97.0% completeness using compleasm in the Eudicots and Embryphyta databases, respectively. Nevertheless, because this previous assembly did not remove duplications from the assembly [43], some assembly errors may exist for gene prediction. For the completed BUSCOs, our assembly revealed 39.7% and 46.8% higher complete and single-copy BUSCOs using BUSCO in the Eudicots and Embryphyta database, while it was 37.4 and 44.0% complete and single-copy BUSCOs in Eudicots and Embryphyta for the previously reported assembly. Using compleasm, our assembly was shown to have 47.9 and 54.7% complete and single-copy BUSCOs in the Eudicots and Embryphyta database, while it was 44.54 and 50.9% complete and single-copy BUSCOs in the Eudicots and Embryphyta database for the previously reported assembly. Therefore, our assembly contained a few duplication errors in the assembly for better gene prediction.

Declarations

Availability of data and materials

Raw sequenced reads have been uploaded to the NCBI Sequence Read Archive under accession number SRR27392947 for long whole genome sequencing reads [10], SRR27392979 for short whole genome sequencing reads [11], SRR27392978 for RNA-seq reads [12], and JAYKKT000000000 for the assembled genome [13]. Please further see Table 1 for details and references [32-34,36,38-42] of the results of the annotations submitted to figshare.

Authors' contributions

L-F. W., W.-G. Z., E-P. Y. and Z.-F. W. collected the samples and wrote the manuscript. W.-G. Z., E-P. Y. and Z.-F. W. generated the sequencing data. L-F. W., H.-L. C. and Z.-F. W. conceived and designed the project. Z.-F. W. analyzed the data. All of the authors have read and approved the final version of this manuscript.

Funding

The study is supported by Guangdong Provincial Forestry Bureau Project — Project of Constructing Model Site for Small and Miniature Wetlands Protection and Restoration in Huadu; Planning of the Provincial Plant Ex Situ Protection System and National Key Protected Plant Ex Situ Protection and Propagation. Key-Area Research and Development Program of Guangdong Province (2022B1111230001) and its **sub-project** (2022B1111230001-2-5). Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden (2023B1212060046).

Acknowledgements

We thank the reviewers for their time, expertise, and helpful suggestions to improve our manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Drzymulska D. On the history of *Brasenia* Schreb. in the European Pleistocene. Veget Hist Archaeobot. 2018;27:527–34. https://doi.org/10.1007/s00334-017-0652-9.
2. Xie C, Li J, Pan F, Fu J, Zhou W, Lu S, Li P, Zhou C. Environmental factors influencing mucilage accumulation of the endangered *Brasenia schreberi* in China. Sci Rep. 2018;8:17955. https://doi.org/10.1038/s41598-018-36448-3.
3. Kim C, Jung J, Na HR, Kim S, Li W, Kadono Y, Shin H, Choi H-K. Population genetic structure of the endangered *Brasenia schreberi* in South Korea based on nuclear ribosomal spacer and chloroplast DNA sequences. J Plant Biol. 2012;55:81–91. https://doi.org/10.1007/s12374-011-9193-4.
4. Xiao H, Cai X, Fan Y, Luo A. Antioxidant activity of water-soluble polysaccharides from *Brasenia schreberi*. Phcog Mag. 2016;12:193–7. https://doi.org/10.4103/0973-1296.186343.
5. Li J, Yi C, Zhang C, Pan F, Xie C, Zhou W, Zhou C. Effects of light quality on leaf growth and photosynthetic fluorescence of *Brasenia schreberi* seedlings. Heliyon. 2021;7:e06082. https://doi.org/10.1016/j.heliyon.2021.e06082.
6. Thompson KA, Sora DM, Cross KS, Germain JMS, Cottenie K. Mucilage reduces leaf herbivory in Schreber's watershield. Brasenia schreberi J F Gmel (Cabombaceae) Bot. 2014;92:412–6. https://doi.org/10.1139/cjb-2013-0296.
7. Elakovich SD. Allelopathic aquatic plants for aquatic weed management. Biol Plant. 1989;31:479–86. https://doi.org/10.1007/BF02876221.
8. Lu B, Shi T, Chen J. Chromosome-level genome assembly of watershield (*Brasenia schreberi*). Sci Data. 2023;10:467. https://doi.org/10.1038/s41597-023-02380-z.
9. Li Z, Gichira AW, Wang Q, Chen J. Genetic diversity and population structure of the endangered basal angiosperm *Brasenia schreberi* (Cabombaceae) in China. PeerJ. 2018;6:e5296. https://doi.org/10.7717/peerj.5296.
10. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plant. NCBI Seq Read Archive. 2024. https://identifiers.org/ncbi/insdc.sra:SRR27392947.
11. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. NCBI Seq Read Archive. 2024.

https://identifiers.org/ncbi/insdc.sra:SRR27392979.

12. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. NCBI Seq Read Archive. 2024. https://identifiers.org/ncbi/insdc.sra:SRR27392978.
13. Joshi NA, Fass JN, Sickle. A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. (2011) Available at: https://github.com/najoshi/sickle. Accessed 24 Aug 2022.
14. Chikhi R, Medvedev P. Informed and automated *k*-mer size selection for genome assembly. Bioinformatics. 2014;30:31–7. https://doi.org/10.1093/bioinformatics/btt310.
15. Porchop v0.2.4. Available at: https://github.com/rrwick/Porechop. Accessed 4 November 2022.
16. Ontbc v1.1. : Pipeline for oxford nanopore barcoding. Available at: https://github.com/FlyPythons/ontbc. Accessed 26 Aug 2022.
17. NextDenovo v2. 3.1: Fast and accurate de novo assembler for long reads. Available at: https://github.com/Nextomics/NextDenovo. Accessed 24 January 2023.
18. Pseudohaploid. Create a pseudohaploid assembly from a partially resolved diploid assembly. Available at:https://github.com/schatzlab/pseudohaploid. Accessed 26 January 2023.
19. Guan DF, McCarthy SA, Wood J, Howe K, Wang YD. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8. https://doi.org/10.1093/bioinformatics/btaa025.
20. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46. https://doi.org/10.1101/gr.214270.116.
21. Aury JM, Istace B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom Bioinform. 2021;3(2):lqab034. https://doi.org/10.1093/nargab/lqab034.
22. Wick RR, Holt KE, Polypolish. Short-read polishing of long-read bacterial genome assemblies. PLoS Comput Biol. 2022;18(1):e1009802. https://doi.org/10.1371/journal.pcbi.1009802.
23. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. NCBI Nucleotide. 2024. https://identifiers.org/nucleotide:JAYKKT000000000.
24. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. Methods Mol Biol. 2019;1962:227–45. https://doi.org/10.1007/978-1-4939-9173-0_14.
25. Huang N, Li H. compleasm: a faster and more accurate reimplementation of BUSCO, Bioinformatics. 2023; 39:btad595. https://doi.org/10.1093/bioinformatics/btad595.
26. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25036598.

27. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25036601. Figshare.
28. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25036607.
29. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25036613.
30. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinform. 2015;16:227. https://doi.org/10.1186/s12859-015-0654-5.
31. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275. https://doi.org/10.1186/s13059-019-1905-y.
32. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25092815.v1.
33. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25092836.v1.
34. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25092983.v1.
35. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. 2023. https://doi.org/10.1101/2023.06.10.544449.
36. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of Brasenia schreberi, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093304.v1.
37. Palmer J, Funannotate. Eukaryotic Genome Annotation Pipeline. Available at:https://github.com/nextgenusfs/funannotate. Accessed 20 Sep 2022.
38. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093376.v2.
39. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093382.v1.

40. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093385.v1.
41. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093394.v1.
42. Wu L-F, Zhu W-G, Yu E-P, Cao H-L, Wang Z-F. Draft genome of *Brasenia schreberi*, a worldwide distributed and endangered aquatic plants. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25093397.v1.
43. Kong W, Wang Y, Zhang S, Yu J, Zhang X. Recent advances in assembly of plant complesx genomes. Genom Proteom Bioinf. 2023;21(3):427–39. https://doi.org/10.1016/j.gpb.2023.04.004.

Tables

Table 1
Overview of all data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Raw long whole genome sequencing reads	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR27392947 [10]
Data file 2	Raw short whole genome sequencing reads	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR27392979 [11]
Data file 3	Raw RNA reads of leaf tissues	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR27392978 [12]
Data file 4	Assembled genome	Fasta file (.fasta)	NCBI Nucleotide, https://identifiers.org/nucleotide:JAYKKT000000000 [23]
Data file 5	BUSCO assessment of the assembly using Eudicots database	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25036598 [26]
Data file 6	BUSCO assessment of the assembly using Embryphyta database	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25036601 [27]
Data file 7	Compleasm assessment of the assembly using Eudicots database	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25036607 [28]
Data file 8	Compleasm assessment of the assembly using Embryphyta database	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25036613 [29]
Data file 8	Repetitive sequences predicted by RED	Text file (.bed)	Figshare, https://doi.org/10.6084/m9.figshare.25092815.v1 [32]
Data file 9	Repetitive sequences predicted by EDTA	Gff3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.25092836.v1 [33]
Data file 10	Repetitive sequences combined by RED and EDTA	Text file (.bed)	Figshare, https://doi.org/10.6084/m9.figshare.25092983.v1 [34]

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 11	Table 1 Species with their protein sequences used for gene prediction	Table (.xlsx)	Figshare, https://doi.org/10.6084/m9.figshare.25093304.v1 [36]
Data file 12	Predicted gene	Gff3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.25093376.v2 [38]
Data file 13	Predicted genes - nucleotide sequences	Fasta file (.fasta)	Figshare, https://doi.org/10.6084/m9.figshare.25093382.v1 [39]
Data file 14	Predicted genes - translated sequences	Fasta file (.fasta)	Figshare, https://doi.org/10.6084/m9.figshare.25093385.v1 [40]
Data file 15	Gene annotation using GO, Pfam, interPro, UniProt, dbCAN, MEROPS and SignalP databases	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25093394.v1 [41]
Data file 16	Gene annotation from eggNOG-mapper analysis	Text (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.25093397.v1 [42]