BMC Genomic Data

DATA NOTE Open Access

The chromosome-level genome of water hyacinth (*Eichhornia crassipes*)

Zhihao Qian[1], Jingshan Yang[1,2], Zhizhong Li[1*] and Jinming Chen[1*]

Abstract

Objectives Water hyacinth (*Eichhornia crassipes*) is one of the most notorious invasive aquatic plants in the world and is known to cause significant ecological and socioeconomic impacts. Here, we reported a high-quality chromosome-level genome for water hyacinth, which will be a valuable reference for future investigations of its invasion.

Data description A chromosome-level genome for water hyacinth was constructed by combing MGI short-reads sequencing, PacBio HiFi (High-fidelity) sequencing, and Hi-C sequencing, which resulted in ca. 1132.2 Mb in size the contig and scaffold N50 length of 18.76 Mb and 69.84 Mb, respectively. A total of 1024.36 Mb (90.47%) of the assembled sequences were anchored to 16 pseudochromosomes, dividing into subgenome A (468.72 Mb in size) and subgenome B (555.64 Mb in size). A total of 57,683 protein-coding genes were predicted, including 25,445 protein-coding genes for subgenome A and 27,992 protein-coding genes for subgenome B. Furthermore, the LAI and QV scores of the water hyacinth genome were 12.32 and 48.91, respectively.

Keywords *Eichhornia crassipes*, Chromosome-level genome, Genome annotation

Objective

The genus *Eichhornia* (Pontederiaceae) has seven species, all of which are potentially invasive [1]. However, the *Eichhornia* species vary in their global distribution and invasiveness. *E. crassipes* is the most invasive among these species. *E. crassipes*, also known as water hyacinth, because of its beautiful violet flowers make it a popular floating aquatic ornamental. It is native to South America but has naturalized in more than 50 countries on five continents [2]. The species is often considered the most notorious aquatic weed and has been listed as one of the 100 most dangerous invasive species in the International Union for Conservation of Nature (IUCN) [3]. Several characteristics of water hyacinth likely contribute to its successful invasion of aquatic habitats, such as its rapid growth rate, ease of propagation and the high mobility of its free-floating life form [4, 5]. These characteristics produce a large amount of biomass covering the water surface, not only choking waterways and hindering transport, but also causing violent changes in the plant and animal communities of the freshwater environment [6, 7]. Most studies on water hyacinth and its closely related species have focused on the reproductive ecology and genetics of tristyly [8–10], but little is known about its mechanism of explosive invasion.

With the rapid development of sequencing technology and bioinformatics, high-quality reference genomes have been successfully applied to help us understand the molecular mechanisms and processes of biological invasions and reveal the molecular genetic mechanisms

*Correspondence:
Zhizhong Li
lizhizhong@wbgcas.cn
Jinming Chen
jmchen@wbgcas.cn
[1]Aquatic Plant Research Center, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

of heterostyly [10–12]. For example, the high-quality genome of the malignant weed *Pistia stratiotes* revealed that the expansion of NBS-LRR gene families corresponding to disease resistance might contribute to its rapid invasion [11]. Recently, the genome assembly of *E. crassipes* and *E. paniculate* have been published [10, 13, 14]. Among these, only Bisht et al. [14] have explored the pathways contributing to its invasiveness and translational potential based on *E. crassipes* genome, but the genome they reported was not assembled at the chromosome level. Here, our high-quality chromosome-level assembly of *E. crassipes* will provide valuable resources for further exploring the molecular mechanisms of its invasion and other biological characteristics.

Data description

The sample of tetraploid water hyacinth (2n = 4x = 32) [15] was sampled from the Wuhan Botanical Garden, Chinese Academy of Sciences, Hubei, China (30° 33′ N, 114° 24′ E). The young-fresh leaves were collected for genome sequencing. Seven tissues (leaves, roots, flowers, petiole, stolon, pistils, and stamens) were collected for RNA sequencing. Here, three sequencing libraries were constructed for genome sequencing: (1) an MGI short-reads library was prepared and sequenced on a DNBSEQ-T7 platform, generating 119.48 Gb raw data total (Data file 1); (2) a PacBio SMRT library sequenced on a PacBio Sequel II platform with the circular consensus sequencing (CCS) mode. Approximately 37.01 Gb HiFi sequences with an average length of 15.4 kb were generated (Data file 2); (3) The Hi-C library was sequenced on the Illumina HiSeq 2500 platform with paired-end 150 bp reads, generating 122 Gb raw data total. Then constructed seven RNA-seq libraries sequenced on the NovaSeq 6000 platform with the PE150 mode (Data file 4).

First, the genome size of water hyacinth was estimated ca. 1130 Mb using flow cytometry conducted on a BD AccuriTMC6 flow cytometer (BD Biosciences, San Jose, CA, USA), *Nelumbo nucifera* (genome size = 808 Mb) was used as reference (Data file 5). Then, the *k-mer* analysis was utilized to evaluate the genome characterization based on 119.48 Gb MGI data using Jellyfish v2.1.3 [16] and GCE v1.0.0 (https://github.com/fanagislab/GCE), which indicated an estimated genome 1184 Mb in size (Data file 5). PacBio HiFi reads were *de novo* assembled into contigs using Hifiasm v0.16.0 [17] with default parameters. Then, the redundancy of the assembled contigs was removed using the Purge_dups v1.2.5 [18] at default settings, generating a primary genome assembly of size 1132.2 Mb with a contig N50 of 18.76 Mb. To create primary scaffolds, we utilized the 3D-DNA pipeline [19] with default settings. The resulting assembly was then visualized and manually refined according to the heatmap of chromosome interactions using Juicebox v1.8.8 [20]. Finally, the chromosome-level assembly was generated by anchoring 115 contigs to 16 pseudochromosomes with two sets of subgenome assemblies, subgenome A and subgenome B, resulting in a 90.47% anchoring rate (Data file 6). The subgenome A and subgenome B were 468.72 Mb and 555.64 Mb in length, with scaffolds N50 reaching 62.66 Mb and 76.89 Mb, respectively (Data file 5).

Here, four methods were used to assess the accuracy and integrity of genome assembly: (1) approximately 99.44% of MGI short reads can be mapped to the reference genome using BWA v0.7.17 [21]; (2) the Benchmarking Universal Single-Copy Orthologs (BUSCOs) score of genome were calculated by BUSCO v5.6.1 [22] with the embryophyta_odb10 database; (3) the genome assembly was evaluated using the LTR Assembly Index (LAI) [23]; and (4) the consensus quality values (QV) of the genome assemblies were assessed using Merqury v1.3 [24] based on MGI short-reads. The complete BUSCOs of subgenome A and subgenome B was 80.9% and 87.5%, respectively. Subgenome A contained 74.0% single-copy, 6.9% duplicated, 1.2% fragmented, and 17.9% missing BUSCOs, while subgenome B contained 78.9% single-copy, 8.6% duplicated, 1.2% fragmented, and 11.3% missing BUSCOs (Data file 5). Also, the LAI and QV values were 12.32 and 48.91, respectively, indicating the high accuracy and completeness of the water hyacinth genome assembly. In addition, we utilized SubPhaser v1.2.6 [25] for subgenome phasing and verified the accuracy of each subgenome's phasing by manually inspecting the results. Our results indicated that the subgenomes identified by SubPhaser were highly consistent with our phased subgenome assemblies (Data file 5).

The repeat library of the water hyacinth assembly was constructed using RepeatModeler v2.0.1 [26] at default settings. Moreover, the repeat library was adopted to scan the genome assemblies with RepeatMasker v4.1.1 (http://www.repeatmasker.org) to identify repetitive sequences in assembly. We identified 649.82 Mb (57.39%) of repetitive regions in the water hyacinth genome. A combination of ab initio, homology, and RNA-seq-based strategies were used to predict protein-coding genes (PCGs) in the water hyacinth genome. Finally, all gene-predicted results were integrated using EVidenceModeler v1.1.1 [27], then PASA [28] was performed to update the result of EVidenceModeler. For more details on genome annotations, please see Data file 5. A total of 57,683 PCGs were predicted, with 25,445 and 27,992 PCGs in subgenome A and subgenome B, respectively (Data files 7–9). The predicted PCGs were functionally annotated using BLAST v2.9.0 (E-value = 1E-05) [29] against the public protein databases, namely the KEGG [30], GO [31], NR (https://ftp.ncbi.nlm.nih.gov/), Swiss-Prot (http://www.expasy.ch/sprot), InterPro [32], and KOG (https://ftp.ncbi.

Table 1 Overview of all data files/data sets

Labe	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Raw short MGI sequencing reads	Fasta file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR25056966 [33]
Data file 2	Raw long HiFi sequencing reads	Fasta file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR25056964 [34]
Data file 3	Raw Hi-C sequencing reads	Fasta file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR25056965 [35]
Data file 4	Raw RNA-seq reads for seven tissues	Fasta file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR25056967 https://identifiers.org/ncbi/insdc.sra:SRR25056968 https://identifiers.org/ncbi/insdc.sra:SRR25056969 https://identifiers.org/ncbi/insdc.sra:SRR25056970 https://identifiers.org/ncbi/insdc.sra:SRR25056971 https://identifiers.org/ncbi/insdc.sra:SRR25056972 https://identifiers.org/ncbi/insdc.sra:SRR25056973 [36]
Data file 5	Supplementary of the genome	pdf file (.pdf)	Figshare, https://doi.org/10.6084/m9.figshare.23651262 [37]
Data file 6	Assembled genome	Fasta file (.fasta)	NCBI GenBank, https://identifiers.org/ncbi/insdc.gca:GCA_030549335.1 [38]
Data file 7	Predicted gene	Gff3 file (.gff)	Figshare, https://doi.org/10.6084/m9.figshare.23635401 [39]
Data file 8	Predicted gene-CDS	CDS file (.cds)	Figshare, https://doi.org/10.6084/m9.figshare.23635401 [39]
Data file 9	Predicted gene-Protein	Protein file (.pep)	Figshare, https://doi.org/10.6084/m9.figshare.23635401 [39]
Data file 10	Gene annotation using KEGG, GO, InterPro, Swiss-Prot, NR, and KOG databases	Annotation file (.html)	Figshare, https://doi.org/10.6084/m9.figshare.23635401 [39]

nih.gov/pub/COG/KOG/). A total of 56,332 PCGs can be annotated function (Data file 10) (Table 1).

Limitations

Although the average BUSCOs score of the two subgenome sets in the current version (84.2%) of water hyacinth was slightly lower than that of the previous version (87.4%), the higher LAI (current: 12.32, previous: 11.78) and QV (current: 48.91, previous: 42.0) values indicated improved reliability of the genome assembly [13]. Howerer, there are still 71 gaps in the the assembly presented here. In the future, water hyacinth genome can be enhanced to T2T (telomere-to-telomere) level by combining Oxford Nanopore Technologies ultra-long reads, which will generate rich genomic information. Also, the annotation of the genome can be further improved based on full-length transcriptomes.

Acknowledgements
Not applicable.

Authors' contributions
J.C. and Z.L. conceived the idea, supervised the work, and revised the manuscript. Z.Q. prepared the plant materials and wrote the original draft manuscript. Z.Q. and J.Y. analyzed the data. All authors have read and approved the final manuscript.

Funding
This work was supported by the Biological Resources Program, CAS (No. KFJ-BRP-007-009) and the Special Research Assistant Project, Chinese Academy of Sciences (No. E2291M01).

Data availability
All raw sequencing outputs were deposited in the NCBI Sequence Read Archive (SRA) under Bioproject accession PRJNA987371, including SRR25056966 for MGI data, SRR25056964 for HiFi data, SRR25056965 for Hi-C data, and SRR25056967- SRR25056973 for seven tissues RNA-seq data. The assembled genome sequences were deposited into NCBI GenBank under accession number GCA_030549335.1. The genome annotation files, including protein-coding gene structure, protein sequences, and functional annotation files have been deposited into the Figshare database (https://doi.org/10.6084/m9.figshare.23635401). The genome assembly and annotation data are also publicly available in China National GeneBank DataBase (CNGBdb) under accession number CNA0069368 (https://ftp.cngb.org/pub/CNSA/data5/CNP0004498/CNS0830885/CNA0069368/).

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 December 2024 / Accepted: 3 April 2025
Published online: 09 April 2025

References

1. Pellegrini MOO, Horn CN, Almeida RF. Total evidence phylogeny of Pontederiaceae (Commelinales) sheds light on the necessity of its recircumscription and synopsis of *Pontederia* L. Phytokeys. 2018;108:25–83. https://doi.org/10.3897/phytokeys.108.27652.
2. Zhang YY, Zhang DY, Barrett SCH. Genetic uniformity characterizes the invasive spread of water hyacinth (*Eichhornia crassipes*), a clonal aquatic plant. Mol Ecol. 2010;19:1774–86. https://doi.org/10.1111/j.1365-294X.2010.04609.x.
3. Global Invasive Species Database, Species profile. *Eichhornia crassipes.* 2023. Downloaded from http://www.iucngisd.org/gisd/species.php?sc=70.
4. Téllez TR, López EMDR, Granado GL, Pérez EA, López RM, Guzmán JMS. The water hyacinth, *Eichhornia crassipes*: an invasive plant in the Guadiana river basin (Spain). Aquat Invasions. 2008;3(1):42–53. https://doi.org/10.3391/ai.2008.3.1.8.
5. da Cunha NL, Xue HR, Wright SI, Barrett SCH. Genetic variation and clonal diversity in floating aquatic plants: comparative genomic analysis of water

hyacinth species in their native range. Mol Ecol. 2022;31:5307–25. https://doi.org/10.1111/mec.16664.
6. Villamagna AM, Murphy BR. Ecological and socio-economic impacts of invasive water hyacinth (*Eichhornia crassipes*): a review. Freshw Biol. 2010;55:282–98. https://doi.org/10.1111/j.1365-2427.2009.02294.x.
7. Schultz R, Dibble E. Effects of invasive macrophytes on freshwater fish and macroinvertebrate communities: the role of invasive plant traits. Hydrobiologia. 2012;684:1–14. https://doi.org/10.1007/s10750-011-0978-8.
8. Liu WL, Wang YF, Chen Q, Yu SX. Pollination of invasive *Eichhornia crassipes* (Pontederiaceae) by the introduced honeybee (*Apis mellifera* L.) in South China. Plant Syst Evol. 2013;299:817–25. https://doi.org/10.1007/s00606-013-0764-3.
9. Arunkumar R, Maddison TI, Barrett SCH, Wright SI. Recent mating-system evolution in *Eichhornia* is accompanied by cis-regulatory divergence. New Phytol. 2016;211:697–707. https://doi.org/10.1111/nph.13918.
10. Arunkumar R, Wang W, Wright SI, Barrett SCH. The genetic architecture of tristyly and its breakdown to self-fertilization. Mol Ecol. 2017;26:752–65. https://doi.org/10.1111/mec.13946.
11. Qian ZH, Li Y, Yang JS, Shi T, Li ZZ, Chen JM. The chromosome-level genome of a free-floating aquatic weed *Pistia stratiotes* provides insights into its rapid invasion. Mol Ecol Resour. 2022;22:2732–43. https://doi.org/10.1111/1755-0998.13653.
12. Bieker VC, et al. Uncovering the genomic basis of an extraordinary plant invasion. Sci Adv. 2022;8:eabo5115. https://doi.org/10.1126/sciadv.abo511.
13. Huang YJ, Guo LB, Xie LJ, Shang NM, Wu DY, Ye CY, et al. A reference genome of commelinales provides insights into the commelinids evolution and global spread of water hyacinth (*Pontederia crassipes*). GigaScience. 2024;13:giae006. https://doi.org/10.1093/gigascience/giae006.
14. Bisht MS, Singh M, Chakraborty A, Sharma VK. Genome of the most noxious weed water hyacinth (*Eichhornia crassipes*) provides insights into plant invasiveness and its translational potential. iScience. 2024;27(9):110698. https://doi.org/10.1016/j.isci.2024.110698.
15. Isa H, Egbuche KC, Malgwi MM, Tukur NA. Cytological studies in *Eichhornia crassipes*. (Mart) Solms Amer J Plant Physiol. 2013;8(2):50–62. https://doi.org/10.3923/ajpp.2013.50.62.
16. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70. https://doi.org/10.1093/bioinformatics/btr011.
17. Cheng HY, Concepcion GT, Feng XW, Zhang HW, Li H. Haplotype-resolved de Novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5. https://doi.org/10.1038/s41592-020-01056-5.
18. Guan DF, McCarthy SA, Wood J, Howe K, Wang YD, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8. https://doi.org/10.1093/bioinformatics/btaa025.
19. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a One-Click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8. https://doi.org/10.1016/j.cels.2016.07.002.
20. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De Novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5. https://doi.org/10.1126/science.aal3327.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
22. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199.
23. Ou SJ, Chen JF, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic Acids Res. 2018;46:e126–126. https://doi.org/10.1093/nar/gky730.
24. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:1–27. https://doi.org/10.1186/s13059-020-02134-9.
25. Jia KH, Wang ZX, Wang L, Li GY, Zhang W, Wang XL, et al. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific *k*-mers. New Phytol. 2022;235:801–9. https://doi.org/10.1111/nph.18173.
26. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117.
27. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol. 2008;9:1–22. https://doi.org/10.1186/gb-2008-9-1-r7.
28. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66. https://doi.org/10.1093/nar/gkg770.
29. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32(suppl2):W20–5. https://doi.org/10.1093/nar/gkh435.
30. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H. Kanehisa. M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 1999;27:29–34. https://doi.org/10.1093/nar/28.1.27.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9. https://doi.org/10.1038/75556.
32. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:211–5. https://doi.org/10.1093/nar/gkn785.
33. Qian ZH, Yang JS, Li ZZ, Chen JM. The chromosome-level genome of water hyacinth (*Eichhornia crassipes*). NCBI Sequence Read Archive. 2023. https://identifiers.org/ncbi/insdc.sra:SRR25056966.
34. Qian ZH, Yang JS, Li ZZ, Chen JM. The chromosome-level genome of water hyacinth (*Eichhornia crassipes*). NCBI Sequence Read Archive. 2023. https://identifiers.org/ncbi/insdc.sra:SRR25056964.
35. Qian ZH, Yang JS, Li ZZ, Chen JM. The chromosome-level genome of water hyacinth (*Eichhornia crassipes*). NCBI Sequence Read Archive. 2023. https://identifiers.org/ncbi/insdc.sra:SRR25056965.
36. Qian ZH, Yang JS, Li ZZ, Chen JM. The chromosome-level genome of water hyacinth (*Eichhornia crassipes*). NCBI Sequence Read Archive. 2023. https://identifiers.org/ncbi/insdc.sra:SRR25056973.
37. Qian ZH. Supplementary of *Eichhornia crassipes* genome. Figshare. 2023. https://doi.org/10.6084/m9.figshare.23635401.
38. Qian ZH, Yang JS, Li ZZ, Chen JM. The chromosome-level genome of water hyacinth (*Eichhornia crassipes*). NCBI GenBank. 2023. https://identifiers.org/ncbi/insdc.gca:GCA_030549335.1.
39. Qian ZH. Annotation of *Eichhornia crassipes* genome. Figshare. 2023. https://doi.org/10.6084/m9.figshare.23635401.

Publisher's Note