BMC Genomic Data

DATA NOTE Open Access

Haplotype-resolved and gap-free genome of a floating aquatic plant from the Oryzeae tribe, *Hygroryza aristata*

Li-Yao Yang[1,4,5], Li-Kun Huang[6], Jin-Bin Lin[1,2,3], Cun-Jing Xu[1,4,5], Wei-Qi Tang[1,2,3*] and Bi-Guang Huang[4,5*]

Abstract

Objectives *Hygroryza aristata*, an aquatic plant native to Southeast Asia, shows a high degree of adaptability to aquatic environments. *H. aristata*, which belongs to the *Oryzeae* tribe and is closely related to rice (*Oryza sativa*), holds potential for crop improvement, particularly in flood tolerance. This study aimed to sequence and assemble the genome of *H. aristata*.

Data description We assembled the genome of *H. aristata* using 31.91 Gb of Pacific Biosciences (PacBio) High-fidelity (HiFi) data and 22.36 Gb of ultra long Oxford Nanopore Technology (ONT) data, resulting in two gap-free haplotype genomes, hap1 (349.74 Mb) and hap2 (347.98 Mb), each with 12 chromosomes and 23 telomeres. The continuity of chromosomes was supported by High-throughput chromosome conformation capture (Hi-C) data. The assemblies demonstrated high completeness, with > 99.8% of coverage rates, 98.4% of Benchmarking Universal Single-Copy Orthologs (BUSCO) scores, and > 11.0 of Long Terminal Repeat Assembly Index (LAI) scores per haplotype. RNA sequencing (RNA-seq) data (176.06 Gb) of six tissues was generated for genome annotation, identifying 39,139 and 38,746 protein-coding genes in hap1 and hap2, respectively.

Keywords *Hygroryza aristata*, Haplotype-resolved, Gap-free, Genome assembly, Genome annotation, PacBio HiFi, Ultra long ONT, Hi-C, RNA-seq

*Correspondence:
Wei-Qi Tang
twq@jointgene.com
Bi-Guang Huang
hbg1989@163.com
[1]College of Geography and Oceanography, Minjiang University, Fuzhou, Fujian 350108, China
[2]Marine and Agricultural Biotechnology Laboratory, Fuzhou Institute of Oceanography, Minjiang University, Fuzhou, Fujian 350108, China
[3]Marine Biotechnology Center, Institute of Oceanography, Minjiang University, Fuzhou, Fujian 350108, China
[4]Key Laboratory of Ministry of Education for Genetics, Breeding and Multiple Utilization of Crops, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China
[5]Fujian Key Laboratory of Crop Breeding by Design, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China
[6]Xiamen Jointgene Biotechnology Co., Ltd, Xiamen, Fujian 361026, China

Objective

Hygroryza aristata (Retz.) Nees ex Wight & Arn. is a perennial aquatic plant native to the wet tropical regions, distributed across Southern China and Southeast Asia [1, 2]. This species has been designated as a second-level nationally protected wild plant in China since 2021. *H. aristata* floats on the water surface, absorbing nutrients directly from the water and rapidly growing and spreading through stoloniferous growth. Morphologically, *H. aristata* produces adventitious roots and branches at the stem nodes, and its leaf sheaths turn purple and inflate, which contributes its floating (Data file 1) [3]. This species shows a high degree of adaptability to aquatic environments.

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Taxonomically and evolutionarily, *H. aristata* (*Hygroryza*, *Zizaniinae*, *Oryzeae*, *Poaceae*) is closely related to rice (*Oryza sativa*), and they both belong to the *Oryzeae* tribe [4]. Unlike most grasses, which favor dry ecosystems, species of the *Oryzeae* tribe, including *H. aristata*, tend to adapt to aquatic environments [5, 6]. Especially, the floating trait is particularly rare within the *Poaceae* family.

Although the chloroplast and mitochondrial genomes of *H. aristata* have been previously reported [7, 8], its nuclear genome has not yet been sequenced. This study presents the complete genomes of the two haplotypes of *H. aristata*. These genomic resources are pivotal for elucidating the evolutionary adaptation to aquatic environments within the *Oryzeae* tribe, and will contribute to the gene resource development for rice, such as those related to flood tolerance.

Data description

Genomic DNA and RNA samples of *H. aristata* were extracted from plants clonally propagated from a single individual. Long-read sequencing of Pacific Biosciences (PacBio) High-fidelity (HiFi) and ultra-long (UL) Oxford Nanopore Technology (ONT) (read lengths > 100 kb), and short-read sequencing of High-throughput chromosome conformation capture (Hi-C), whole-genome sequencing (WGS), and RNA sequencing (RNA-seq), were performed (Data file 2) [3].

For genome assembly, 31.91 Gb of PacBio HiFi and 22.36 Gb of UL-ONT sequencing data sets were utilized (Data file 2; Data set 1) [3, 9]. Assembly was conducted using HiFiAsm (v0.20.0-r639) [10] under HiFi + UL-ONT mode with the following parameters: -l 3 -r 5 -a 6 -n 10 --ctg-n 10 -w 63 -k 63. Chromosome IDs and strand directions were determined by aligning the assemblies to the rice genome [11]. The resulting assemblies of unphased two haplotypes, designated as hap1 and hap2, were obtained (Data set 2, 3) [12, 13].

Both hap1 and hap2 are complete genomes, each comprising 12 chromosomes with genome sizes of 349.74 Mb and 347.98 Mb, respectively (Data file 3) [3]. Notably, both haplotypes are gap-free (Data file 3) [3]. Telomere detection using Seqtk telo (v1.4-r122) [14] displayed that each haplotype contains 23 telomeres (Data file 3) [3]. In conclusion, this study presents a haplotype-resolved and gap-free genome assembly.

To assess genome, genomic reads mapping, Benchmarking Universal Single-Copy Orthologs (BUSCO) [15] and Long Terminal Repeat Assembly Index (LAI) [16] methods were employed. Hi-C reads (83.06 Gb; Data file 2; Data set 1) [3, 9] were aligned to hap1 and hap2 using Juicer (v2.0) [17], respectively. Short-read

Table 1 Overview of 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	Figure 1. Plant morphology of *H. aristata*	Image file (.png)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data file 2	Table 1. Summary of sequencing data of *H. aristata*	Word file (.docx)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data file 3	Table 2. Chromosomes information of *H. aristata*	Excel file (.xlsx)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data file 4	Figure 2. Hi-C contact heatmaps of two haplotypes of *H. aristata*	Image file (.pdf)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data file 5	Table 3. Protein sequence sets of 16 genomes of 15 species in *Poaceae* used for genome annotation	Excel file (.xlsx)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data set 1	*Hygroryza aristata* genome and transcriptome sequencing data	BAM file (.bam) and FASTQ files (.fastq)	NGDC Genome Sequence Archive, https://ngdc.cncb.ac.cn/gsa/browse/CRA019829 [9]
Data set 2	hari_v2.2_hap1 (genome assembly and annotation of hap1)	FASTA file (.fasta) and GFF file (.gff)	NGDC Genome warehouse, https://ngdc.cncb.ac.cn/gwh/Assembly/86182/show [12]
Data set 3	hari_v2.2_hap2 (genome assembly and annotation of hap2)	FASTA file (.fasta) and GFF file (.gff)	NGDC Genome warehouse, https://ngdc.cncb.ac.cn/gwh/Assembly/86205/show [13]
Data set 4	hari_v2.2_hap1.EDTA.TEanno.gff3.gz (repeat annotation of hap1)	GFF file (.gff)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data set 5	hari_v2.2_hap2.EDTA.TEanno.gff3.gz (repeat annotation of hap2)	GFF file (.gff)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data set 6	hari_v2.2_hap1_anno_v1.1.total.annatation.xlsx (integrated coding gene function annotations of hap1)	Excel file (.xlsx)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]
Data set 7	hari_v2.2_hap2_anno_v1.1.total.annatation.xlsx (integrated coding gene function annotations of hap2)	Excel file (.xlsx)	Zenodo, https://doi.org/10.5281/zenodo.14105124 [3]

WGS data (45.49 Gb; Data file 2; Data set 1) [3, 9] and long reads were aligned to the diploid genome assembly (hap1 + hap2, ~700 Mb) using bwa mem (v0.7.18-r1243) [18] and minimap2 (v2.28-r1209) [19], respectively.

The continuity of hap1 and hap2 chromosomes was supported by the Hi-C contact results (Data file 4) [3]. The alignments of WGS data achieved a 99.864% coverage rate with a mean depth of 63.1×, while PacBio HiFi and UL-ONT alignments reached 99.996% and 99.995% coverage rates with mean depths of 44× and 31.9×, respectively. The complete BUSCO scores were 98.4% for both hap1 and hap2 assemblies, and the LAI scores were 11.98 for hap1 and 11.19 for hap2.

For genome annotation, 176.06 Gb of RNA-seq data (Data file 2; Data set 1) [3, 9] and proteins from 16 genomes of 15 species in *Poaceae* (Data file 5) [3], were employed. Gene structure prediction was conducted using Braker3 (v3.0.8) [20], combining homology prediction, transcriptional evidences and ab initio prediction. Repeat annotation was conducted using EDTA (v2.0.1) [21]. The number of protein-coding genes of hap1 and hap2 were 39,139 and 38,746, respectively (Data set 2, 3) [12, 13]. The content of repeat sequences of hap1 and hap2 were 49.08% and 49.57%, respectively (Data set 4, 5) [3]. Finally, we annotated genes, including non-redundant protein database (NR), Universal Protein Knowledgebase (UniProt), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Protein families database (PFAM), as well as other related public databases (Data set 6, 7) [3] (Table 1).

Limitations

Despite the high quality of the current genome assembly of *H. aristata*, there is still room for enhancement to achieve an error-free level. This can be accomplished by integrating various assembly outcomes from different assemblers, conducting fine revisions, and leveraging cytological experimental data, among other approaches.

The genome annotation of *H. aristata* can be further improved. For examples, non-coding RNAs have not been identified in the current version of the annotation. Future updates to the annotation are planned, which will incorporate data from third-generation full-length transcriptome sequencing of additional tissue samples to enhance the annotation.

Abbreviations

PacBio	Pacific Biosciences
HiFi	High-fidelity
ONT	Oxford Nanopore Technology
Hi-C	High-throughput chromosome conformation capture
WGS	Whole-genome sequencing
RNA-seq	RNA sequencing
BUSCO	Benchmarking Universal Single-Copy Orthologs
LAI	Long Terminal Repeat (LTR) Assembly Index

Acknowledgements

Not applicable.

Authors' contributions

W-Q T and B-G H conceived and designed the project. L-Y Y, B-G H and W-Q T collected the samples. L-Y Y and W-Q T generated sequences and assembled the genome. L-K H and W-Q T performed data management. L-K H annotated the genome. J-B L and C-J X performed quality assessments. L-Y Y, B-G H and W-Q T wrote the manuscript. All authors approved the manuscript.

Funding

The study is supported by grant from Natural Science Foundation of Fujian Province of China (2024J0113).

Data availability

The genome and transcriptome sequencing data used for assembly and annotation have been deposited in the Genome Sequence Archive (GSA) of National Genomics Data Center (NGDC) under the accession number: CRA019829. The genome assembly and annotation results of two haplotypes (hap1 and hap2) have been deposited in the Genome Warehouse (GWH) of NGDC under the accession numbers: GWHFGLW00000000.1 and GWHFGNQ00000000.1, respectively. The repeat annotation and protein-coding gene function annotation results have been deposited in the Zenodo data repository (https://doi.org/10.5281/zenodo.14105124).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 November 2024 / Accepted: 19 March 2025
Published online: 28 March 2025

References

1. Flora of China. 2024. http://efloras.org/florataxon.aspx?flora_id=2&taxon_id=200025546.
2. The World Flora Online. 2024. https://www.worldfloraonline.org/taxon/wfo-0000875137.
3. Tang W. Haplotype-resolved and gap-free genome of a floating aquatic plant from the Oryzeae tribe, Hygroryza aristata (hari_v2.2) [Data set]. Zenodo; 2024. https://doi.org/10.5281/zenodo.14105124.
4. Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Teisher JK, Clark LG, et al. A worldwide phylogenetic classification of the *Poaceae* (*Gramineae*) II: an update and a comparison of two 2015 classifications: phylogenetic classification of the grasses II. J Syt Evol. 2017;55:259–90. https://doi.org/10.1111/jse.12262.
5. Vaughan DA, Morishima H, Kadowaki K. Diversity in the *Oryza* genus. Curr Opin Plant Biol. 2003;6:139–46. https://doi.org/10.1016/s1369-5266(03)00009-8.
6. Zhang T, Huang W, Zhang L, Li D-Z, Qi J, Ma H. Phylogenomic profiles of whole-genome duplications in *Poaceae* and landscape of differential duplicate retention and losses among major *Poaceae* lineages. Nat Commun. 2024;15:3305. https://doi.org/10.1038/s41467-024-47428-9.
7. Wang H, Wu Z, Fan G, Zheng G, Gu C, Tembrock LR, et al. Characterization of the complete chloroplast genome of *Hygroryza aristata* (Retz.) Nees ex Wight & Arn. (Zizaniinae, Poaceae). Mitochondr DNA B Resour. 2021;6:1949–50. https://doi.org/10.1080/23802359.2021.1935352.
8. Wang H, Wu Z, Li T, Zhao J. Highly active repeat-mediated recombination in the mitogenome of the aquatic grass *Hygroryza aristata*. BMC Plant Biol. 2024;24:644. https://doi.org/10.1186/s12870-024-05331-x.
9. Tang W. Hygroryza aristata genome and transcriptome sequencing data. NGDC Genome Sequence Archive. 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA019829.

10. Cheng H, Concepcion GT, Feng X, Zhang H, 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.
11. Shang L, He W, Wang T, Yang Y, Xu Q, Zhao X, et al. A complete assembly of the rice Nipponbare reference genome. Mol Plant. 2023;16:1232–6. https://doi.org/10.1016/j.molp.2023.08.003.
12. Tang W. hari_v2.2_hap1. NGDC Genome warehouse. 2024. https://ngdc.cncb.ac.cn/gwh/Assembly/86182/show.
13. Tang W. hari_v2.2_hap2. NGDC Genome warehouse. 2024. https://ngdc.cncb.ac.cn/gwh/Assembly/86205/show.
14. Github. 2024. https://github.com/lh3/seqtk.
15. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351.
16. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 2018;46:e126. https://doi.org/10.1093/nar/gky730.
17. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. 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.
18. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95. https://doi.org/10.1093/bioinformatics/btp698.
19. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021:btab705. https://doi.org/10.1093/bioinformatics/btab705.
20. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77. https://doi.org/10.1101/gr.278090.123.
21. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. 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.

Publisher's Note

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