nature portfolio

Corresponding author(s): Bin Han

Last updated by author(s): Nov 3, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data. Sequencing platforms used to generate the raw data are listed as followed: Pacific Biosciences Sequel II, PromethION, Illumina HiSeq 4000, Illumina NovaSeq.
Data analysis	We used publicly available and appropriately cited software in the Methods. No commercial software and code were used in this study. Software are listed as follows: 1. raw reads assembly: Hifiasm (version 0.16.0), Necat (version 20200803), NextDenovo (version 2.1), Racon (version 1.0.0), NextPolish (version 1.0.2), BWA (version 0.7.17), chromap (version 0.2.6), samtools (version 1.20), YaHS (version 1.1), juice_tools (version 1.19.02), JBAT (version 1.9.8). 2. pseudo-chromosomes construction: MUMMER (version 4.0.0beta2), ALLMAPS. 3. genome assessment: BUSCO (version 5.2.2), LTR_retriever (version 2.9.0), Inspector (version 1.0.1), QUAST (version 5.0.1). 4. identification of centromeres and telomeres: ClustalW (version 2.1), HMMER (version 3.1b2), MAFFT (version 7.490), IQ-TREE (version 1.6.12). 5. gene annotation and expression levels calculation: FGENESH+ (version 3.1.1), SNAP (version 2006-07-28), GeneMark-ES (version 4.68_lic) and AUGUSTUS (version 3.3.2), GenomeThreader (version 1.7.1), HISAT2 (version 2.0.5), StringTie (version 2.0), Trinity (version 2.12.0), PASA (version 2.0.1), EVidenceModeler (version 1.1.1), InterProScan (version 5.56-89.0), fastp (version 0.23.0), Salmon (version 1.6.0). 6. identification of allelic genes and differential expression allelic genes:GeneTribe(version 1.2.0),BLASTN (version 2.9.0+), GMAP (version 2021-05-27) 7. identification of resistance gene analogs: RGAugury, MCscan (Python version). 8. annotation of transposable elements: EDTA (version 2.1.0), panEDTA, RepeatMasker (version 4.1.2), LTR_retriever (version 2.9.0), R package clusterProfiler (version 4.6.2).

9. SV calling: pbmm2 (version 1.4.0), PBSV (version 2.6.2), minimap2 (version 2.21-r1071), CuteSV (version 1.0.11), SVIM-asm (version 1.0.7), SURVIVOR (version 1.0.6), SyRI (version 1.4).
10. SNP calling: longshot (version 0.4.1), minimap2 (version 2.21-r1071), MUMMER package (version 4.0.0beta2), GATK (version 4.1.4.0), SnpEff (version 55.0).
10. pan-genome construction: BLASTN (version 2.2.18), minigraph (version 0.19-r551), Cactus (version 2.2.1), vg giraffe (version 1.43.0), DeepVariants (version 1.6.1), GLnexus (version 1.4.1-0-g68e25e5).
11. evolutionary analysis: HomBlocks, DIAMOND (version 2.0.15), OrthoFinder (version 2.5.4), IQ-TREE (version 2.2.0.3), PLINK (version 1.90b6.9 64-bit), EMMAX (version beta-07Mar2010), PHYLIP (version 3.66), iTOL, VCFTools (version 0.1.16), archetypal-analysis, PopLDdecay (version 3.42), R (version 4.1.3), Python (veresion 3.8.10), BEDTools (version 2.30.0), RectChr (version 1.36).
12. demographic history inference: MSMC2 (version 2.1.4), seqbility (version 20091110).
13. gene flow analysis: TreeMix (version 1.13), R package OptM (version 0.1.6), AdmixTools (version 7.0.2), genomics_general toolkit.
14. haplotype analysis: GMAP (version 2021-05-27), MAFFT (version 7.490), R package geneHapR (version 1.1.9).
15. Distribution map: GeoPandas (version 0.14.4), numpy (version 1.26.4), pandas (version 2.2.2), matplotlib (3.8.4)
16. All the analysis scripts used in this study are available at Github (https://github.com/dongling-hub/Wild-rice-Pangenome-Project) and Zenodo (https://doi.org/10.5281/zenodo.14881729) repository.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All raw sequencing data, transcriptome data and Hi-C data generated in this study have been deposited in the European Nucleotide Archive (ENA) under the BioProject accession number PRJEB73710. The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center (NGDC), Beijing Institute of Genomics, Chinese Academy of Sciences/ China National Center for Bioinformation with the BioProject accession number PRJCA024131. All assemblies with annotations, variant VCF files and graph pangenome files are available at Figshare (https://doi.org/10.25452/figshare.plus.25697817) and the RicePandb database (http://ricepandb.ncgr.ac.cn). The embryophyta_odb10 database, used for genome completeness assessment, was download from the https://busco-data.ezlab.org/v4/data/lineages/. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not applicable
Reporting on race, ethnicity, or other socially relevant groupings	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	PacBio HiFi: 133 samples; Nanopore: 16 samples; Hi-C: 30 samples; Illumina: 696 samples. We built the pan-genome based on 145 representative accessions in our study. We selected this number of populations because they sufficiently cover the genetic and geographic distribution of both Oryza rufipogon and Oryza sativa. These materials are evenly distributed within an evolutionary tree that includes a larger set of approximately 1500 samples.

Data exclusions	No data was excluded.
Replication	The study primarily focuses on genome assembly and evolutionary analyses, relying on genetic data and computational models rather than experimental protocols that would necessitate replication.
Randomization	A randomized complete block design was used in planting.
Blinding	All accessions were only labeled by numbers when planting and data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☐	☒	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Hazards	*Please describe the agents/technologies/information that may pose a threat, including any agents subject to oversight for dual use research of concern.*

For examples of agents subject to oversight, see the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Precautions and benefits

Biosecurity precautions	*Describe the precautions that were taken during the design and conduct of this research, or will be required in the communication and application of the research, to minimise biosecurity risks. These may include bio-containment facilities, changes to the study design/ methodology or redaction of details from the manuscript.*
Biosecurity oversight	*Describe any evaluations and oversight of biosecurity risks of this work that you have received from people or organizations outside of your immediate team.*
Benefits	*Describe the benefits that application or use of this work could bring, including benefits that may mitigate risks to public health, national security, or the health of crops, livestock or the environment.*
Communication benefits	*Describe whether the benefits of communicating this information outweigh the risks, and if so, how.*

Plants

Seed stocks	We selected a total of 149 rice varieties according to the phylogenetic relationships and geographic distribution from a previously reported population, which were cultivated in Lingshui County, Hainan.
Novel plant genotypes	We only collected wild accessions and cultivars in this study. No novel plant were used.
Authentication	All samples were from the China National Rice Research Institute in Hangzhou, China, and the National Institute of Genetics in Mishima, Japan.