


# An improved haplotype resolved genome reveals more rice genes

Muhammad Abdullah<sup>1,2</sup>, Agnelo Furtado<sup>1</sup>, Ardashir Kharabian Masouleh<sup>1</sup>, Pauline Okemo<sup>1,2</sup> and

Robert J. Henry<sup>1,2\*</sup> 

<sup>1</sup> Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, QLD, 4067, Australia

<sup>2</sup> ARC Centre for Plant Success in Nature and Agriculture, University of Queensland, Brisbane, QLD, 4067, Australia

\* Corresponding author, E-mail: [robert.henry@uq.edu.au](mailto:robert.henry@uq.edu.au)

## Abstract

The rice reference genome (*Oryza sativa* ssp. *japonica* cv. Nipponbare) has been an important resource in plant science. We now report an improved and haplotype resolved genome sequence based upon more accurate sequencing technology. This improved assembly includes regions missing in earlier genome sequences and the annotation of more than 3,000 new genes due to greater sequence accuracy. This phased genome will be a useful resource for rice research.

**Citation:** Abdullah M, Furtado A, Masouleh AK, Okemo P, Henry RJ. 2024. An improved haplotype resolved genome reveals more rice genes. *Tropical Plants* 3: e009 <https://doi.org/10.48130/tp-0024-0007>

## Introduction

Nipponbare is a japonica rice cultivar that has been widely used as the standard reference genotype for rice<sup>[1]</sup>. The rice (Nipponbare) genome was one of the first crop genomes to be sequenced more than 20 years ago<sup>[2]</sup>. The 1<sup>st</sup> sequence of the rice genome was completed in 2002 and was a major milestone in the field of plant genomics by the International Rice Genome Sequencing Project, 2005<sup>[3]</sup>. These international collaborative efforts provided the first genome of a crop plant. The Nipponbare genome assembly contained gaps, primarily due to repetitive DNA sequences. In 2005, these gaps were estimated to be approximately 18.1 Mb in total, with the majority originating from centromeres and telomere regions. Sequencing technological advancements and ongoing research efforts, have improved the rice genome sequence over time<sup>[4,5]</sup>. Thorough efforts were made to improve the quality of the Nipponbare reference genome assembly in 2013, resulting in greatly enhanced accuracy of cDNA sequences and gene annotation, while it remained incomplete<sup>[5]</sup>. In the human genome, recent significant strides have been made in assembling and characterization the previously unexplored 8% of the human genome, especially including telomere sequences<sup>[6]</sup>.

Reference genome assemblies often contain gaps, especially regions with repetitive sequences, termed the 'dark side' of the genome<sup>[7,8]</sup>. New sequence technology allows improved assembly quality, with less gaps, leading to a more complete and accurate representation of the genome. The achievement of a higher quality and more complete reference genome will provide new insights into genomics and breeding, supporting pan-genome studies and genome wide association studies<sup>[9]</sup>. Recently many other *Oryza* genomes have been sequenced and assembled, including indica and wild rice species<sup>[9–11]</sup>. Most recently the Nipponbare genome sequence gaps and telomere sequence were addressed<sup>[12,13]</sup>. Despite these advancements, a fully haplotype resolved assembly has not been reported. In this study, we have used PacBio HiFi reads to produce a more

accurate genome sequence assembly. The novel genome assembly is not only almost 11.3 Mb longer than the IRGSP-1.0 reference but also exhibits improvements in all chromosomes (Fig. 1), including the addition of telomeric regions in all chromosomes (T2T), with the addition of fully resolved haplotypes (haplotype 1 and haplotype 2, telomere-to-telomere-T2T) (Tables 1, 2 & Supplemental Tables S1–S10).

Comparative analysis of annotations of the new genome (UQ Nipponbare) and the IRGSP-1.0 reference revealed the presence of 3,050 additional genes, for which more than 95% had supporting transcript evidence (Supplemental Fig. S1).

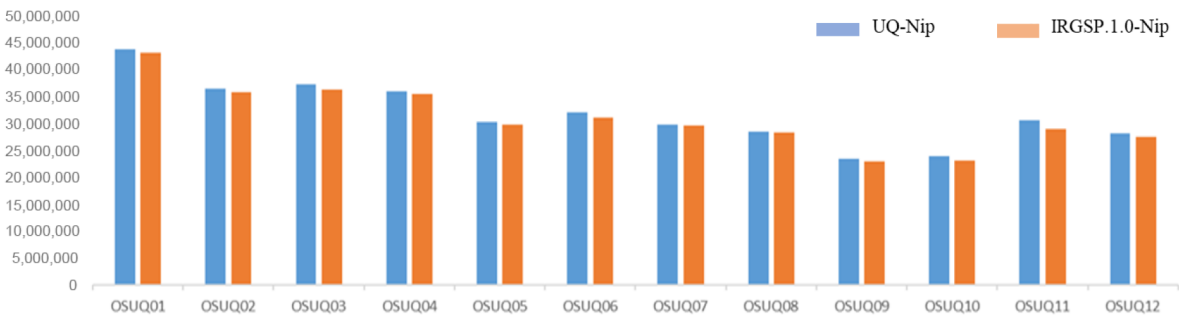
These findings underscore the potential of new sequencing technologies to significantly augment reference genomes, potentially leading to more comprehensive genetic information. These results also suggest that applying advanced sequencing technologies to other established genomes may yield similar benefits, potentially enhancing our knowledge of these species. This study highlights the continuous evolution of genomics and underscore the importance of staying at the forefront of sequencing technologies for the accurate representation of complex genomes.

PacBio HiFi reads and Hi-C reads were used to generate a contig assembly with Hifiasm<sup>[14]</sup> producing a haplotype phased assembly. The contig level assembly produced single contigs for nine chromosomes, while the remaining three chromosomes were each covered by two contigs each. Hi-C data were employed to hierarchically cluster the assembled contigs into 12 pseudo-chromosomes, by using the YaHS scaffolding tool<sup>[15]</sup>. The T2T assembly had a single scaffold for each of the 12 pseudo-chromosomes and was larger in size than the corresponding IRGSP.10 genome (Fig. 1). The results of BUSCO analysis showed that the collapsed assembly covered 99.3% of the universal single copy genes with an N50 of 30.7 Mb (Supplemental Tables S1–S3). The UQ Nipponbare collapsed genome assembly is larger in size compared to the IRGSP.1.0 Nipponbare reference genome assembly. In Fig. 2, additional

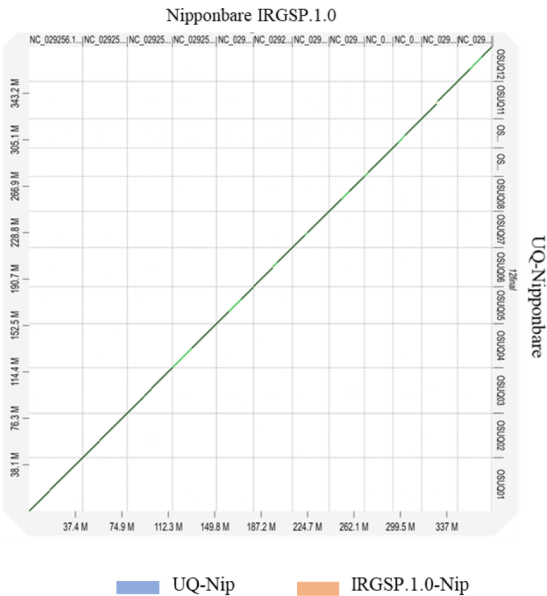
a

UQ_Nipponbare			Nipponbare-IRGSP.1.0		
#Chr	UQ_NIP	Telomere	#Chr	Nip-IRGSP.1.0	Telomeres
OSUQ01	43,960,277	2	chr1	43,270,923	0
OSUQ02	36,506,049	2	chr2	35,937,250	0
OSUQ03	37,404,130	2	chr3	36,413,819	0
OSUQ04	36,083,220	2	chr4	35,502,694	0
OSUQ05	30,417,698	2	chr5	29,958,434	0
OSUQ06	32,132,640	2	chr6	31,248,787	0
OSUQ07	29,813,588	2	chr7	29,697,621	0
OSUQ08	28,607,546	2	chr8	28,443,022	0
OSUQ09	23,461,744	1	chr9	23,012,720	0
OSUQ10	23,948,751	2	chr10	23,207,287	0
OSUQ11	30,712,252	2	chr11	29,021,106	0
OSUQ12	28,269,131	2	chr12	27,531,856	0

b



c



**Fig. 1** Comparison of the UQ Nipponbare genome with previously published reference genome assembly IRGSP.1.0 Nipponbare. (a), (b) For UQ Nipponbare all the chromosome sizes are larger and most include telomeres as compared to IRGSP.1.0 Nipponbare. (c) Whole genome dot plot of UQ Nipponbare genome vs IRGSP.1.0 Nipponbare.

**Table 1.** Statistics for the UQ Nipponbare haplotype resolved genome assembly.

	UQ_Nip-collapsed	UQ_Nip-Hap1	UQ_Nip-Hap2
Total assembly size	381,317,026	379,234,557	348,265,595
Complete BUSCOs (%)	99.30%	98.90%	94.90%
Total scaffold number	12	12	12
Scaffold N50	30,712,252	30,691,512	29,307,860
Scaffold L50	6	6	6
Largest scaffold	43,960,277	43,881,444	37,312,016
GC content (%)	44	44	43

non-aligning regions of each chromosome in the UQ Nipponbare collapsed assembly are highlighted, along with the structural variants in the comparison of the previously published IRGSP1.0 Nipponbare reference genome assembly.

For the two phased haplotypes (T2T) of UQ Nipponbare; haplotype 1 covered 98.9% of the single copy orthologs with an N50 of 30.6 Mb, whilst haplotype 2 covered 94.9% single copy orthologs with an N50 of 29.3 Mb (Supplemental Tables S4–S9 & Supplemental Figs S2 & S3). The haplotype 1 chromosomes were larger than the haplotype 2 chromosomes (Fig. 3). This first haplotype resolved Nipponbare genome incorporated 3,050 new genes compared to IRGSP-1.0, and is expected to be a valuable and significant resource for rice researchers, and these additional genes had a wide range of functions (Supplemental Table S11). Of these additional genes, 58 genes

fell in new regions that were missing in the IRGSP genome, but most genes were in regions that were not new due to the improved accuracy of sequencing.

Methods

DNA extraction and sequencing

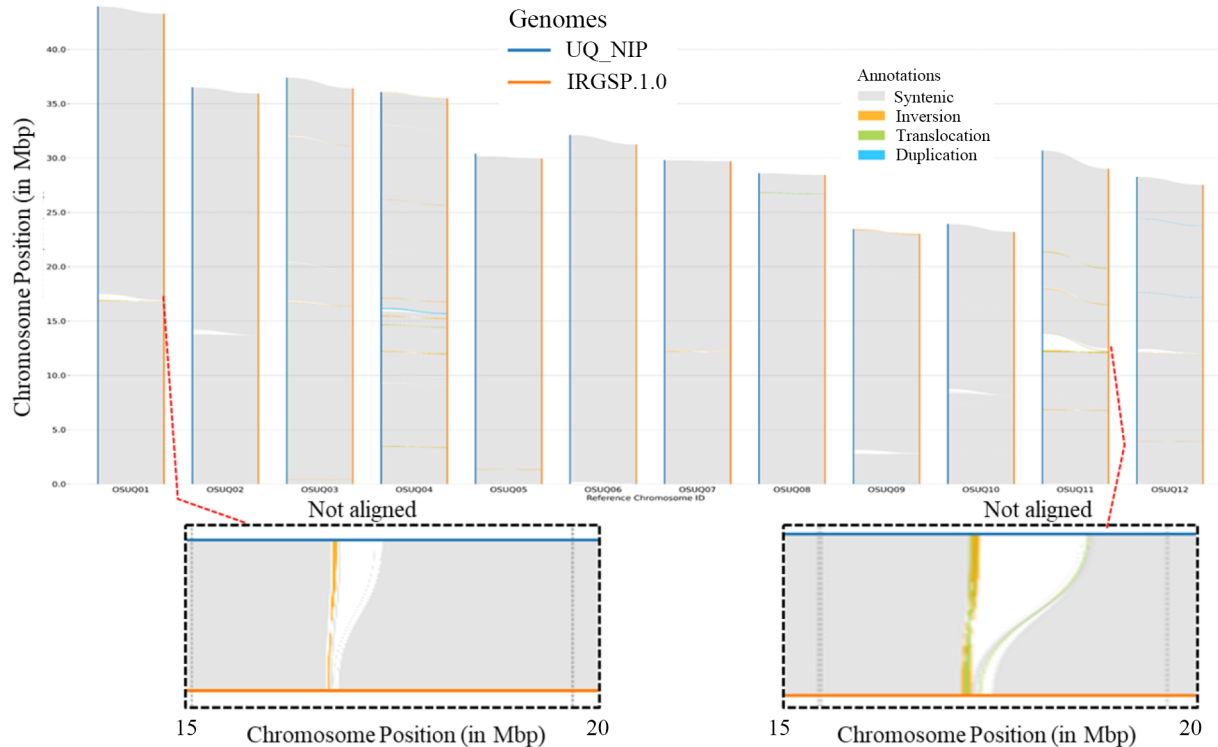
The CTAB method<sup>[19]</sup> was used to extract DNA from young leaves of a rice (*Oryza sativa* cv Nipponbare) plant grown in a glasshouse at the University of Queensland (Australia). The high quality DNA extracted was sequenced using a PacBio (Pacific Biosciences) Sequel II platform to produce HiFi sequences.

Genome assembly

Approximately 54.9 Gb of HiFi reads were obtained. HiC reads (59.6 Gb) were downloaded from the NCBI Sequence Read Archive database (SRR6470741). *De novo* haplotype-resolved assembly of these reads was performed using hifiasm with parameters '--write-ec --write-paf -l0'<sup>[14]</sup>. The contig assemblies were scaffolded using the YaHS tool<sup>[15]</sup>. QUAST and BUSCO were used to evaluate the quality and completeness of a genome assembly<sup>[16,17]</sup>. A telomere identification toolkit (tidk) was used to search for tandem repeats of the telomeric sequence 'TTAGGG' and 'TAAACCC' and the exact location in the Nipponbare collapse, haplotype-1 and haplotype-2 assembly (<https://github.com/tolkit/telomeric-identifier>).

**Table 2.** UQ Nipponbare haplotype resolved genome assembly chromosomes sizes and telomere numbers.

UQ_Nip-Collapsed-Assembly			UQ_Nip-Hap1			UQ_Nip-Hap2		
Chr	Size	Telomere	Chr	Size	Telomeres	Chr	Size	Telomeres
OSUQ01	43,960,277	2	OSUQ01-hap1-01	43,881,444	2	OSUQ01-hap2-01	37,312,016	2
OSUQ02	36,506,049	2	OSUQ02-hap1-02	36,408,562	2	OSUQ02-hap2-02	33,128,268	2
OSUQ03	37,404,130	2	OSUQ03-hap1-03	37,357,616	2	OSUQ03-hap2-03	35,934,962	2
OSUQ04	36,083,220	2	OSUQ04-hap1-04	35,866,358	2	OSUQ04-hap2-04	33,298,206	2
OSUQ05	30,417,698	2	OSUQ05-hap1-05	30,178,946	2	OSUQ05-hap2-05	24,819,415	2
OSUQ06	32,132,640	2	OSUQ06-hap1-06	32,049,832	2	OSUQ06-hap2-06	32,067,105	2
OSUQ07	29,813,588	2	OSUQ07-hap1-07	29,708,413	2	OSUQ07-hap2-07	28,849,953	2
OSUQ08	28,607,546	2	OSUQ08-hap1-08	28,566,876	2	OSUQ08-hap2-08	26,254,248	1
OSUQ09	23,461,744	1	OSUQ09-hap1-09	23,221,905		OSUQ09-hap2-09	21,159,188	1
OSUQ10	23,948,751	2	OSUQ10-hap1-10	23,192,984	2	OSUQ10-hap2-10	21,409,976	2
OSUQ11	30,712,252	2	OSUQ11-hap1-11	30,691,512	2	OSUQ11-hap2-11	29,307,860	1
OSUQ12	28,269,131	2	OSUQ12-hap1-12	28,110,109	2	OSUQ12-hap2-12	24,724,398	2

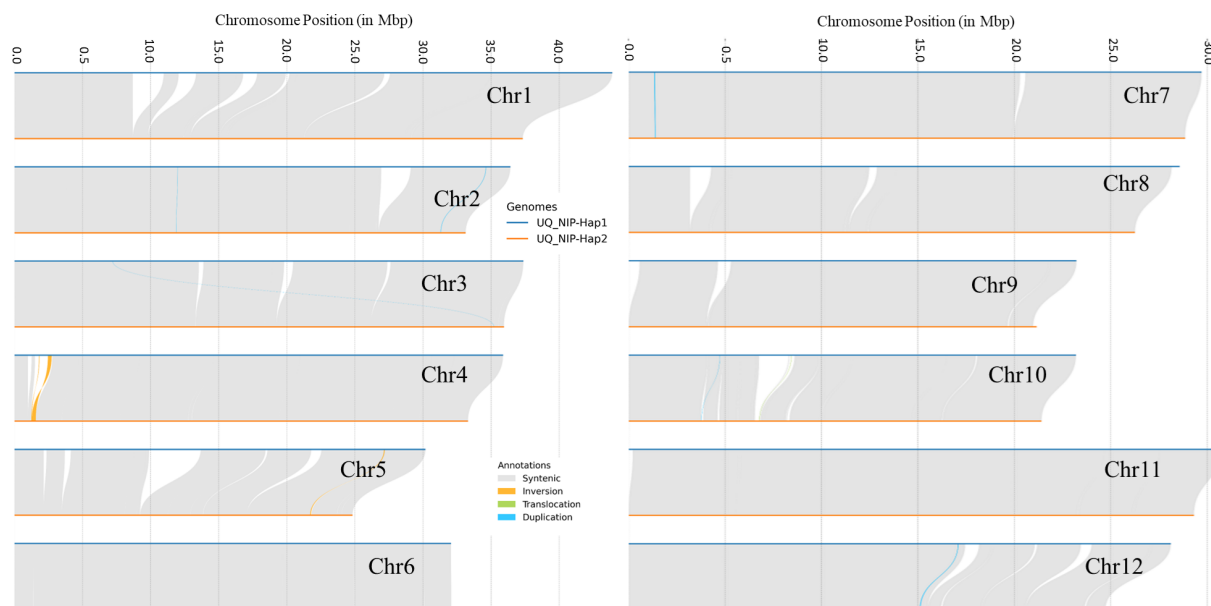


**Fig. 2** Sequence collinearity and structural variants, including inversions, translocations, duplications, and non-aligning regions, were analysed between the UQ Nipponbare genome assembly and the IRGSP-1.0-Nipponbare genome assembly. The two assemblies were aligned using minimap2, and the resulting BAM file was indexed with samtools. Detection of structural variations between these two genomes was performed using SyRI<sup>[26–28]</sup>. The non-aligning regions of chromosomes 1 and 11 are highlighted in the bottom section of the figure.

## Genome annotation

Repetitive DNA sequences were obtained from a *Oryza* repeat database<sup>[18]</sup> and used to mask the genome with the Repeatmasker soft masking option<sup>[19]</sup>. Protein sequences of Viridiplantae from OrthoDB v.11<sup>[20]</sup> and RNA-sequencing (RNA-seq) reads from the NCBI Sequence Read Archive database (SRR23560402, SRR23560417, SRR23560416, SRR23560419, SRR23560418, SRR23560409, SRR23107175, SRR23107177, SRR23107178, SRR8051554, SRR7974062, SRR8051550) were obtained. Quality and adapter trimmed RNA-seq reads were aligned to the masked genomes using HISAT2<sup>[21]</sup>. Annotation of protein-coding genes in Nipponbare was conducted using a combination of homology-based prediction, *de novo* prediction, and transcriptome-based prediction methods using

Braker<sup>[22]</sup>. BUSCO was used to assess the genome annotation completeness. The Large Gap Mapping tool (length fraction; 0.9 similarity fraction; 0.9) of CLC was used to identify the new genes with the comparison of IRGSP-1.0 Nipponbare genes (CLC Genomics Workbench 23.0.05, QIAGEN, USA, [www.clcbio.com](http://www.clcbio.com)) and further transcript evidence for these new genes was estimated (Supplemental Table S12). The functional annotation of the identified additional genes was performed using OmicsBox 2.2.4<sup>[23]</sup>. CDS sequences were subjected to a BLASTX analysis with a specific e-value of 1.0E-10 against the non-redundant protein sequences database, utilizing Viridiplantae taxonomy. Subsequently, the CDS sequences were processed through InterProScan, and GO terms were extracted for all matches acquired via the BLAST search,



**Fig. 3** Sequence collinearity and structural variants, including inversions, translocations, duplications, and non-aligning regions, were examined between the UQ Nipponbare haplotype 1 genome assembly and the UQ Nipponbare haplotype 2 genome assembly. The same method used in Fig. 2 was used for this analysis.

employing Gene Ontology mapping with the Blast2GO annotation tool. The annotations generated from InterProScan and Blast2GO were then harmonized by merging the respective GO terms (Supplemental Table S11).

## Author contributions

The authors confirm contribution to the paper as follows: experiment interpretation: Abdullah M; project supervision: Furtado A, Masouleh AK, Henry RJ; data analysis: Abdullah M, Furtado A, Masouleh AK; draft manuscript preparation and data interpretation: Okemo P; study conception: Henry RJ. All authors approved the final version of the manuscript.

## Data availability

The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center<sup>[24,25]</sup>, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number GWHEQBP00000000 that is publicly accessible at <https://ngdc.cncb.ac.cn/gwh>.

## Acknowledgments

This research was supported by the ARC Centre of Excellence for Plant Success in Nature and Agriculture (Grant No. CE200100015).

## Conflict of interest

The authors declare that they have no conflict of interest. Robert J. Henry is the Editorial Board member of *Tropical Plants* who was blinded from reviewing or making decisions on the manuscript. The article was subject to the journal's standard procedures, with peer-review handled independently of this Editorial Board member and the research groups.

**Supplementary information** accompanies this paper at (<https://www.maxapress.com/article/doi/10.48130/tp-0024-0007>)

## Dates

Received 7 December 2023; Revised 20 January 2024; Accepted 5 February 2024; Published online 3 April 2024

## References

- Matsumoto T, Wu J, Itoh T, Numa H, Antonio B, et al. 2016. The Nipponbare genome and the next-generation of rice genomics research in Japan. *Rice* 9:33
- Jackson SA. 2016. Rice: The first crop genome. *Rice* 9:14
- Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* 296:92–100
- International Rice Genome Sequencing P, Sasaki T. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800
- Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, et al. 2022. The complete sequence of a human genome. *Science* 376:44–53
- Zhao T, Duan Z, Genchev GZ, Lu H. 2020. Closing human reference genome gaps: Identifying and characterizing gap-closing sequences. *G3 Genes | Genomes | Genetics* 10:2801–9
- Chen Q, Lan C, Zhao L, Wang J, Chen B, et al. 2017. Recent advances in sequence assembly: principles and applications. *Briefings in Functional Genomics* 16:361–78
- Li F, Han Z, Qiao W, Wang J, Song Y, et al. 2021. High-quality genomes and high-density genetic map facilitate the identification of genes from a weedy rice. *Frontiers in Plant Science* 12:775051
- Brozynska M, Copetti D, Furtado A, Wing RA, Crayn D, et al. 2017. Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice. *Plant Biotechnology Journal* 15:765–74

## Rice genome sequence

11. Li K, Jiang W, Hui Y, Kong M, Feng LY, et al. 2021. Gapless *indica* rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Molecular Plant* 14:1745–56
12. Shang L, He W, Wang T, Yang Y, Xu Q, et al. 2023. A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* 16:1232–36
13. Huang X. 2023. A complete telomere-to-telomere assembly provides new reference genome for rice. *Molecular Plant* 16:1370–72
14. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18:170–75
15. Zhou C, McCarthy SA, Durbin R. 2023. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39:btac808
16. Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34:i142–i150
17. Manni M, Berkeley MR, Seppely M, Zdobnov EM. 2021. BUSCO: Assessing genomic data quality and beyond. *Current Protocols* 1:e323
18. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, et al. 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research* 35:D883–D887
19. Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 25:4.10.1–4.10.14
20. Kuznetsov D, Tegenfeldt F, Manni M, Seppely M, Berkeley M, et al. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* 51:D445–D451
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37:907–15
22. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* 3:lqaa108
23. OmicsBox – Bioinformatics made easy. 2019. BioBam Bioinformatics (version 2.2.4). [www.biobam.com](http://www.biobam.com)
24. Chen M, Ma Y, Wu S, Zheng X, Kang H, et al. 2021. Genome Warehouse: A public repository housing genome-scale data. *Genomics, Proteomics & Bioinformatics* 19:584–89
25. CNCB-NGDC Members and Partners. 2023. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. *Nucleic Acids Research* 51:D18–D28
26. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–100
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–79
28. Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* 20:277



Copyright: © 2024 by the author(s). Published by Maximum Academic Press on behalf of Hainan University. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.