Plant Communications
Correspondence

CellPress
Partner Journal

Complete genome assembly of *japonica* rice variety Zhonghua 11

Dear Editor,

This study reports a 379.33-Mb telomere-to-telomere genome assembly of the *japonica* rice variety Zhonghua 11, generated using PacBio HiFi and Oxford Nanopore Technologies sequencing data. The high quality of this assembly will serve as a valuable resource for the rice research community.

Oryza sativa ssp. *japonica* var. Zhonghua 11 (ZH11) has been the subject of extensive research, primarily because of its high transformation efficiency and strong adaptability (Tie et al., 2012). However, the ZH11 genome assembly is incomplete and still contains 25 gaps (Rice Resource Center, ricerc.sicau.edu.cn; Qin et al., 2021). In this study, we used PacBio HiFi (HiFi) and Oxford Nanopore Technologies (ONT) methods to obtain a 379.33-Mb, telomere-to-telomere (T2T) assembly of the ZH11 genome.

In recent years, rice research has been greatly facilitated by the release of genome sequences for different rice varieties, including the wild allotetraploid rice *Oryza alta* and *O. sativa* ssp. *indica* cultivars such as 9311 (Yu et al., 2002), Zhengshan97, Minghui63 (rice.hzau.edu.cn/rice_rs3), and R498 (mbkbase.org/R498). Genome sequences of *japonica* cultivars such as Suijing18 have also been released (Nie et al., 2017). The availability of T2T genomes, i.e., complete and gapless T2T genome assemblies, has significantly increased our understanding of genomic structure and function. ZH11, a popular rice variety for experimentation that has been the subject of thousands of studies over the last 20 years (Wu et al., 2003), was developed using the pollen culture method (www.ricedata.cn). The current ZH11 genome assembly is 378.9 Mb (IGDBCAS_Zhonghua11_1.0; Jin et al., 2020) or 376.8 Mb (ZH11-IGDBv1; Qin et al., 2021) in length but contains 25 gaps. To assemble and annotate a T2T-ZH11 genome, we obtained 23.02 Gb of HiFi data, 56.7 Gb of Nanopore long-read data (102×), 24.96 Gb of Illumina data (60×), and 42.59 Gb of high-throughput/resolution chromosome conformation capture (Hi-C) data, as well as 1.89 Gb of full-length transcriptome sequencing data (isoform sequencing [Iso-seq]), 101 Gb of next-generation sequencing transcriptome data, and 246 Gb of Bionano data (Supplemental Table 1). Initially, a draft genome comprising 377.78 Mb of DNA sequence and 17 contigs was assembled from ONT and Illumina reads using NextDenovo and Nextpolish (Supplemental Figure 1; Supplemental Table 2). After removal of 2 contigs corresponding to organelle genomes, the genome size was reduced to 377.05 Mb, with 15 contigs anchored to 12 chromosomes; this was defined as the ZH11 1.0 version (Supplemental Table 3). The ZH11 1.0 genome exhibited high completeness and was validated using Bionano data (Supplemental Data 1); its Benchmarking Universal Single-Copy Orthologs (BUSCO) score was 98.47% and its Core Eukaryotic Genes Mapping Approach (CEGMA) score was 93.15% (Supplemental Tables 4–8; Supplemental Figure 2). Nonetheless, it still had 2 gaps (Supplemental Figure 3), leaving 264 649 bp of unassigned sequences.

To fill the 2 remaining gaps in ZH11 1.0, we used a combination of ONT and HiFi technologies with the circular consensus sequencing algorithm in SMRT Link (v.9.0), which generated HiFi data at a sequencing depth of 60× (Figure 1B; Supplemental Table 9). We assembled the contig version of the genome (Supplemental Table 10) and then analyzed it against nucleotide, mitochondrial, and plastid databases to exclude contaminant sequences, leading to the removal of 62 contigs with a total of 3 059 520 bp. The genome was then assembled from ONT and HiFi reads using Hifiasm and scaffolded into 12 chromosomes from 13 contigs using Hi-C data with Lachesis. This left 1 gap, which was subsequently filled with ONT long reads via quarTeT; filling of this gap and extension of the telomeres resulted in a complete T2T-ZH11 genome of 379.33 Mb (Figure 1A).

The completeness and uniformity of the T2T-ZH11 sequence were supported by a 99.64% mapping rate and 99.89% coverage (at least 1× depth) with Illumina reads, 99.97% mapping rate and 99.98% coverage with HiFi reads, and 99.9% coverage with ONT reads (at least 1× depth). BUSCO analysis revealed 99.01% completeness, similar to that of T2T-Nip (99.88%), including 1560 single and 38 duplicated copies of genes in the Embryophyta database (Figure 1C). CEGMA analysis revealed 99.56% completeness of 458 highly conserved core eukaryotic genes (Figure 1D). We used the Telomere Identification toolKit to detect potential centromeric repeats, and alignment to T2T-ZH11 and T2T-Nip through FindTelomeres identified 24 telomeres and 12 centromeres (Figure 1E; Supplemental Table 11). The accuracy of the assembly was further assessed using 246 Gb of Bionano data (Supplemental Data 1 and 2). We found that the position of the centromere on chromosome 9 was shifted significantly compared with its position in T2T-Nip (Supplemental Table 11), likely owing to differences in chromosome length and sequence loss (Supplemental Figure 4). Chromatin immunoprecipitation sequencing (ChIP-seq) experiments using CENH3 antibodies validated the accuracy of the T2T-ZH11 genome assembly and confirmed the centromere regions, demonstrating that the latter regions were well assembled (Supplemental Figure 5). Collinearity analysis of T2T-ZH11 and ZH11 1.0 confirmed that the 2 gaps were filled, resulting in 2.29 Mb of new sequence; the 2 gaps were located at the end of chromosome 5 (chr5; ∼30 Mb) and in the middle of chr11 (∼12 Mb) (Figure 1F). We analyzed the sequences in the 2 gaps and identified an additional gene, *Osa05G046820*, on chr5 in T2T-ZH11, whose expression was confirmed using RNA sequencing (RNA-seq) data (Supplemental Figure 6).

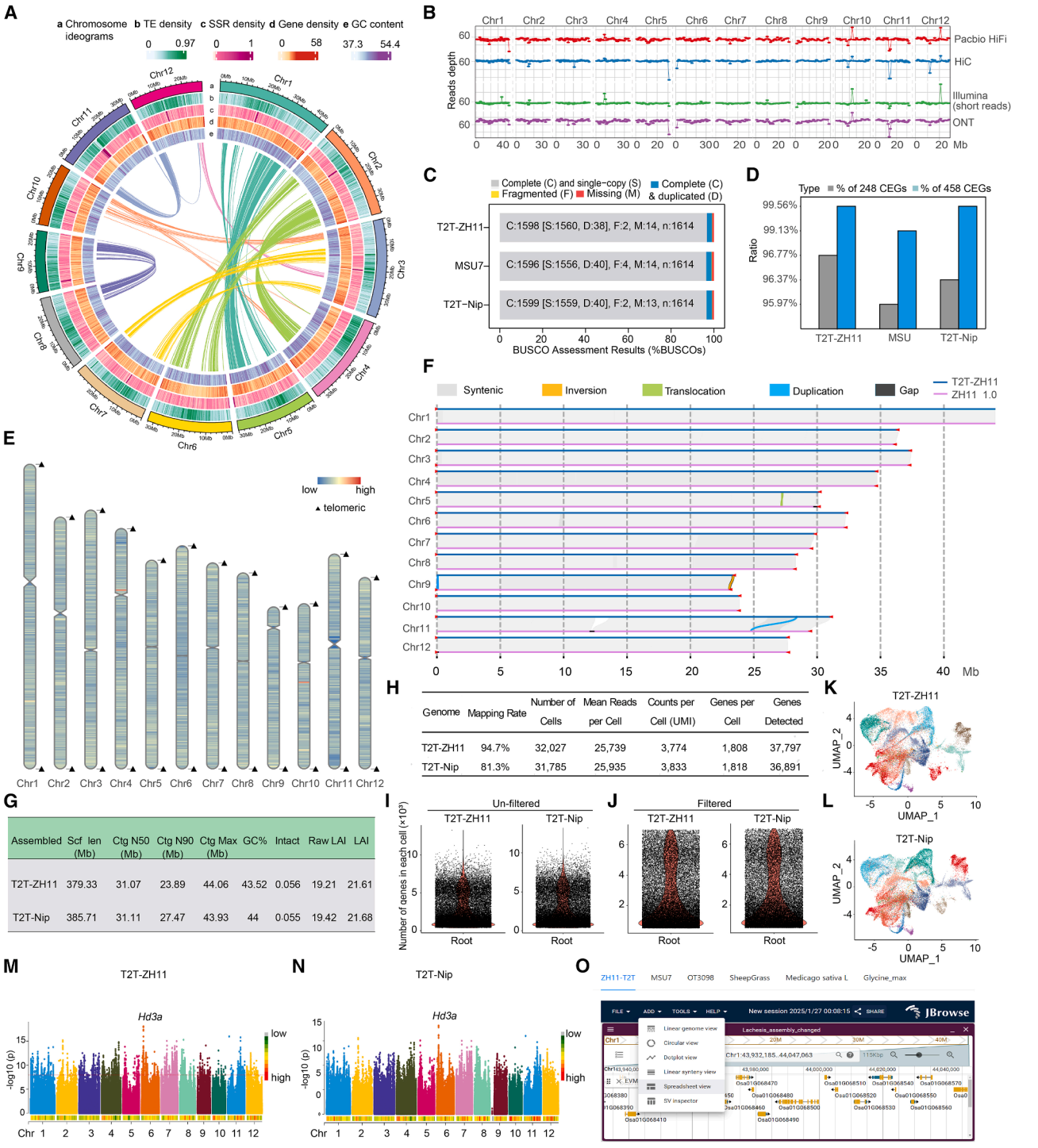

Figure 1. The T2T-ZH11 reference genome has been assembled and validated.

(A) Circos plot and heatmap illustrating genomic features such as transposable elements (TEs), simple sequence repeats (SSRs), genes, and GC content across the T2T-ZH11 genome using 100-kb windows.

(B) The T2T-ZH11 genome sequence was re-aligned using different sequencing technologies (HiFi, ONT, Hi-C, and NGS-Illumina) to assess sequencing depth.

(C and D) Genome evaluation was performed using BUSCO and core eukaryotic genes to compare T2T-ZH11 with T2T-Nip and MSU7.0.

(E) Heatmap showing the positions of telomeres across the 12 chromosomes of T2T-ZH11.

(F) Collinearity analysis of T2T-ZH11 and the ZH11 1.0 reference genome was performed .

(G) Summary of the qualities of the T2T-ZH11 and T2T-Nip genomes, including metrics such as scaffold length (Scf len), contig N50 (Ctg N50) and N90 (Ctg N90), maximum contig (Ctg max), and GC content. LAI values were evaluated using LTR retrotransposons.

(legend continued on next page)

High-accuracy K-mer analysis indicated a consensus quality value of 49.49. We used candidate long terminal repeat (LTR)-RT sequences to accurately identify LTR retrotransposons and calculate an LTR assembly index (LAI) value of 21.61 (Supplemental Table 12; Figure 1G). This value indicated that the T2T-ZH11 assembly was of "gold standard" quality (LAI > 20), comparable to that of the T2T-Nip assembly. Transposable element (TE) predictions identified 410 952 TEs, which accounted for 174.07 Mb of DNA sequence or 45.89% of the genome (Supplemental Table 12).

We annotated T2T-ZH11 and predicted 56 948 genes using Iso-seq data, RNA-seq data, and homologous proteins (Supplemental Table 13). To validate the assembled T2T-ZH11 genome and determine whether it was improved compared with the T2T-Nip assembly, we used 147.56 Gb of published raw single-cell RNA-seq (scRNA-seq) data from roots (Zhang et al., 2021), together with whole-genome sequencing data from 533 varieties. We used Cell Ranger statistics for the scRNA-seq data to remove duplicate UMIs (Unique Molecular Identifiers) and count the number of unique UMIs (≥100) expressed in a single cell; this analysis revealed mapping rates of 94.70% (T2T-ZH11) and 81.3% (T2T-Nip) (Figure 1H), median UMI counts per cell of 25 739 (T2T-ZH11) and 25 935 (T2T-Nip), median genes per cell of 1808 (T2T-ZH11) and 1818 (T2T-Nip), and total gene numbers of 37 797 (T2T-ZH11) and 36 891 (T2T-Nip) (Figure 1I and 1J). Cells were clustered and grouped using the SNN clustering algorithm, then visualized with t-SNE and UMAP (uniform manifold approximation and projection). UMAP plots of the rice root single-cell atlas, aligned with the T2T-ZH11 and T2T-Nip genomes, revealed 14 cell clusters in each (Figure 1I and 1J).

ATAC-seq (SRX25751900 and SRX25751888), RNA-seq (SRX6399012 and SRX6399013), and DNA methylation (SRX6398997, SRX6398998, and SRX6398999) data from the NCBI Sequence Read Archive (SRA) and all annotation RNA-seq data in the ZH11 background were also compared between MSU7.0 and T2T-ZH11 (Supplemental Tables 14 and 15). The results confirmed that using the T2T-ZH11 genome as a reference for alignment improved the mapping rate, sequencing coverage, and number of identified peaks (Supplemental Data 3). In addition, when Diamond alignment was used to filter genes with similarity <90%, 9365 genes with less than 90% identity were found (Supplemental Figure 7) in T2T-ZH11; for example, *Osa01G001700*, whose predicted protein sequence is highly similar to that of ID BAD52506.1 in NCBI, encodes an unknown protein with an amino acid sequence insertion (EVKRKEEREREGEGKGGERGGGL) not found in MSU7.0 (Supplemental Figure 8). Using data from the Rice Genome Annotation Project, we found 13 389 TE-related genes in T2T-ZH11 with >85% similarity to those in MSU7.0 according to alignment (Supplemental Data 4).

We next examined potential associations of heading date and grain weight with over 4 million SNPs from 533 varieties of rice in the RiceVarMap database (RicevarMap2, ricevarmap.ncpgr.cn). To further assess the accuracy of the T2T-ZH11 genome, we performed the genome-wide association study using both the T2T-ZH11 (Figure 1M) and T2T-Nip (Figure 1N) genomes. This analysis identified significant signals for the two traits in candidate genes such as *Hd3a* and *GW8* (Supplemental Figure 9). The same signals were observed in both genomes, providing further support for the reliability of the T2T-ZH11 genome assembly. The divergence times among the 533 varieties and *Oryza rufipogon* (China National Center for Bioinformation, CRR124016) were estimated using a maximum likelihood method, confirming that the divergence of wild rice predates that of cultivated rice (Supplemental Figure 10). Integrated information from the T2T-ZH11 reference genome, together with comparable reference genomes, such as rice T2T-Nip (Shang et al., 2023), rice MSU7.0 (Rice Genome Annotation Project), sheep grass (Li et al., 2023), and *Medicago sativa* L. (Chen et al., 2020), were deposited at the CROPTILLING website (www.croptilling.com/ZH11), where they are now accessible (Figure 1O). Various genome quality metrics, including the BUSCO score, CEGMA score, quality value, and LAI value, demonstrate the high accuracy and completeness of the T2T-ZH11 genome assembly.

Overall, the complete T2T-ZH11 reference genome comprises 379.33 Mb of DNA sequence on 12 chromosomes and has a contig N50 of 31.07 Mb. We filled 25 gaps, increasing the assembly by 2.29 Mb compared with ZH11-IGDBv1, and we predicted 56 948 genes (including 13 028 TE-related genes), 792 tRNAs, 840 rRNAs, and 811 microRNAs. The full genome assembly also revealed the presence of approximately 174.07 Mb of TE sequences and 18.09 Mb of tandem repeat sequences, which represented 45.89% and 4.77% of the complete T2T-ZH11 genome, respectively.

The advent of HiFi and ONT sequencing technologies, coupled with the availability of Hifiasm algorithms, have made it feasible to generate high quality end-to-end genome assemblies. These developments have also led to the production of longer contigs and effectively addressed the issue of fragmented repeats, thus significantly improving the quality of genome assemblies. Here, the integration of Iso-seq technology improved the gene annotations, increasing the number of genes to 56 948, which is as many as that in T2T-Nip (Supplemental Table 16). The quality of the

(H–J) The quality of the T2T-ZH11 genome was evaluated using single-cell RNA sequencing (scRNA-seq) data from rice roots (Zhang et al., 2021). The reads were aligned with the T2T-ZH11 and T2T-Nip genomes. Mapping rates and number of genes per cell were then compared; unfiltered and filtered data are shown in **(I)** and **(J)**, respectively. The quality control thresholds were established as follows: mitochondrial gene content ≤20% (to exclude cells with potential apoptosis/lysis artifacts), minimum UMI count per cell ≥100 (to ensure sufficient transcript capture), and gene detection requirement of ≥10 cells per gene (to remove low-confidence genomic features). UMAP plots of the rice root single-cell atlas, which were aligned with the T2T-ZH11 and T2T-Nip genomes, are shown in **(K** and **L)**, respectively; each dot represents a cell, and different colors indicate different clusters.

(M and N) Manhattan plot showing the results of a genome-wide association study for heading date, identifying significant signals and genes such as *Hd3a* using T2T-ZH11 and T2T-Nip as reference genomes.

(O) Information from the T2T-ZH11 reference genome is shown alongside that from other reference genomes and is available online (www.croptilling.com/ZH11).

T2T-ZH11 genome is also comparable to that of T2T-Nip, as both were assembled using HiFi and ONT data, Hifiasm algorithms, and the same analytical pipelines (Supplemental Table 17). This consistency in sequencing and assembly methods ensured that both genomes meet high standards of quality and accuracy. Compared with previously published genomes (Supplemental Table 18), the T2T-ZH11 genome assembly is substantially better in terms of both continuity and completeness.

DATA AND CODE AVAILABILITY

Raw sequence data for ZH11 1.0 have been deposited in the NCBI SRA database under BioProject ID PRJNA592760. The following sequencing data are available in the NCBI SRA database: Nanopore long reads (SRR10589848), Illumina short reads (SRR10567117), Hi-C reads (SRR10568176 and SRR10568177), and RNA-seq data (SRR10568171, SRR10568172, and SRR10568173). The whole-genome sequence data for T2T-ZH11 have been deposited in the National Genomics Data Center, China National Center for Bioinformation, under PRJCA037115 (CRA027482 and CRA023865) and are publicly accessible at https://ngdc.cncb.ac.cn. ATAC-seq (SRX25751900 and SRX25751888), RNA-seq (SRX6399012 and SRX6399013), and DNA methylation data (SRX6398997, SRX6398998, and SRX6398999) are publicly available at NCBI.

FUNDING

This work was supported by a grant from the Natural Science Foundation of China (32230077), the Innovation Program of the Chinese Academy of Agricultural Sciences (CAAS-CSIAF-202303), and the National Key R&D Program of China (2021YFF1000200).

ACKNOWLEDGMENTS

The authors would like to acknowledge the NCBI and NGDC staff for help with uploading the sequencing data and Biomarker Technologies Co., Ltd. for technical assistance with sequencing and bioinformatics analysis. No conflict of interest is declared.

AUTHOR CONTRIBUTIONS

C.-M.L. and H.H. conceived the project and guided the study. X.-F.Y. prepared the materials as well as the experimental design. X.-F.Y. and Z.S. carried out analyses for validation. X.-F.Y. drafted the manuscript. X.-F.Y., Z.S., and X.X. revised the manuscript.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: March 12, 2025
Revised: April 22, 2025
Accepted: July 17, 2025
Published: July 22, 2025

Xue-Feng Yao[1,5], Zongyi Sun[3], Xintong Xu[1,4], Huihui Li[2,6] and Chun-Ming Liu[1,2,4,5,7,]*

[1]Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[2]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[3]Grandomics Biosciences Co., Ltd., Wuhan 430076, China
[4]School of Advanced Agricultural Sciences, Peking University, Beijing 100091, China
[5]University of Chinese Academy of Sciences, Beijing 100049, China
[6]Nanfan Research Institute, CAAS, Sanya, Hainan 572024, China
[7]The Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing 210095, China
***Correspondence: Chun-Ming Liu (cmliu@ibcas.ac.cn)**
https://doi.org/10.1016/j.xplc.2025.101463

REFERENCES

Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y., Liu, Y., et al. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat. Commun. **11**:2494. https://doi.org/10.1038/s41467-020-16338-x.

Jin, S., Fei, H., Zhu, Z., Luo, Y., Liu, J., Gao, S., Zhang, F., Chen, Y.H., Wang, Y., and Gao, C. (2020). Rationally designed APOBEC3B cytosine base editors with improved specificity. Mol. Cell **79**:728–740.e6. https://doi.org/10.1016/j.molcel.2020.07.005.

Li, T., Tang, S., Li, W., Zhang, S., Wang, J., Pan, D., Lin, Z., Ma, X., Chang, Y., Liu, B., et al. (2023). Genome evolution and initial breeding of the *Triticeae* grass *Leymus chinensis* dominating the Eurasian Steppe. Proc. Natl. Acad. Sci. USA **120**:e2308984120. https://doi.org/10.1073/pnas.2308984120.

Nie, S.J., Liu, Y.Q., Wang, C.C., Gao, S.W., Xu, T.T., Liu, Q., Chang, H.L., Chen, Y.B., Yan, P.C., Peng, W., et al. (2017). Assembly of an early-matured japonica (Geng) rice genome, Suijing18, based on PacBio and Illumina sequencing. Sci. Data **4**:170195. https://doi.org/10.1038/sdata.2017.195.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell **184**:3542–3558.e16. https://doi.org/10.1016/j.cell.2021.04.046.

Shang, L., He, W., Wang, T., Yang, Y., Xu, Q., Zhao, X., Yang, L., Zhang, H., Li, X., Lv, Y., et al. (2023). A complete assembly of the rice Nipponbare reference genome. Mol. Plant **16**:1232–1236. https://doi.org/10.1016/j.molp.2023.08.003.

Tie, W., Zhou, F., Wang, L., Xie, W., Chen, H., Li, X., and Lin, Y. (2012). Reasons for lower transformation efficiency in *indica* rice using *Agrobacterium tumefaciens*-mediated transformation: lessons from transformation assays and genome-wide expression profiling. Plant Mol. Biol. **78**:1–18. https://doi.org/10.1007/s11103-011-9842-5.

Wu, C., Li, X., Yuan, W., Chen, G., Kilian, A., Li, J., Xu, C., Li, X., Zhou, D.X., Wang, S., et al. (2003). Development of enhancer trap lines for functional analysis of the rice genome. Plant J. **35**:418–427.

Yu, J., Hu, S., Wang, J., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296**:79–92. https://doi.org/10.1126/science.1068037.

Zhang, T.Q., Chen, Y., Liu, Y., Lin, W.H., and Wang, J.W. (2021). Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. Nat. Commun. **12**:2053. https://doi.org/10.1038/s41467-021-22352-4.