CellPress
Partner Journal

Molecular Plant
Correspondence

A complete assembly of the rice Nipponbare reference genome

Dear Editor,

In 2005, the current commonly used rice reference genome (*Oryza sativa* ssp. *japonica* cv. Nipponbare) was initially released by the International Rice Genome Sequencing Project (International Rice Genome Sequencing Project, 2005). Thereafter, the reference genome was further updated in 2013 with improved genome assembly (IRGSP-1.0) and gene annotations (MSU7, RAP-DB) (Kawahara et al., 2013; Sakai et al., 2013). In the past 10 years, this reference has been serving as one of the most important genetic resources for subsequent rice functional genomics efforts. As several rice genomes had been assembled into gapless chromosomes with only 2–5 telomeres absent (Li et al., 2021; Song et al., 2021; Zhang et al., 2022), the IRGSP-1.0 and its annotations still performed as the most widely used reference. However, limitations of sequencing technology and intricate genomic organization led to an under-representation of complex regions in this reference, leaving a total of 72 major gaps (including 19 telomeres), 167 minor gaps, and 779 unknown bases (Kawahara et al., 2013), with an estimated length of ~3% of the genome unsolved.

To pursue a complete sequence of this foundational reference genome, we applied a hybrid assembly strategy that integrated Pacbio HiFi and Oxford Nanopore Technology (ONT) ultra-long reads to generate original contigs, which were then scaffolded onto a chromosome-level assembly with the support of the Hi-C dataset. Gap filling and terminal extension were further conducted to resolve the remaining seven gaps and one telomere region within the scaffolds. All gap-closure regions were supported with uniform coverage of ONT reads (Supplemental Figure 1). A large rDNA array was identified beside the telomere of short arm in chromosome 9 with nearly identical repeats of 45S rDNA (Supplemental Figure 2), which was artificially filled with consecutive blocks reflecting their estimated copy number (see supplemental materials and methods). This captured 93.8% of HiFi reads and 93.9% of ONT reads containing 45S rDNA by full-length mapping, but should be treated as model sequences. Following sequence polishing employing the HiFi and Illumina PE (next-generation sequencing [NGS]) reads, we produced a complete assembly of the rice reference genome, T2T-NIP (version AGIS-1.0), within which all 12 centromere and 24 telomere regions were resolved (Figure 1A).

Multiple strategies were applied to evaluate the accuracy and completeness of T2T-NIP. All available primary data—including HiFi, ONT, NGS, and Hi-C—were remapped to T2T-NIP with high mapping rates of >99.6% in all datasets except for ONT reads (93.1%). All reads displayed uniform coverage across the whole genome, except for the Hi-C dataset because of large centromeres and complex regions near two telomeres (Figure 1B). Chromatin immunoprecipitation and sequencing (ChIP-seq) were conducted with the rice CENH3 antibody to identify the location and sequence of functional centromeres in T2T-NIP (Figure 1A, Supplemental Table 1, and Supplemental Figure 3). CentO-enriched regions were also identified by sequence homology to the 155- to 165-bp CentO satellite repeats (Figure 1A and Supplemental Table 1), eight of which showed similar or consistent size with a previous report as determined by fluorescence *in situ* hybridization (Cheng et al., 2002). The consensus accuracy of the whole genome was estimated to be approximately one error per 5 million bases (Q63), which showed much higher sequence accuracy (Supplemental Table 2). For gene content assessment, T2T-NIP captured 99.88% of a BUSCO 1614 gene set (Supplemental Table 3), which was equal to or higher than previously reported gapless rice genomes (Li et al., 2021; Song et al., 2021; Zhang et al., 2022). A total of 1747 ribosomal RNA (rRNA) genes were identified in T2T-NIP, whereas only several hundred were identified in the IRGSP-1.0. A total of 57 359 protein-coding genes and 325 794 repeat elements (51.1%) were identified, both of which represent more than for IRGSP-1.0 (Supplemental Tables 4 and 5). In a model sequence of the 45S rDNA array, 1022 genes were annotated with support of transcriptome data (Supplemental Table 6). Among 314 protein-coding genes annotated in the gap-filling regions excluding this rDNA array, 142 genes were confirmed to be expressed in T2T-NIP and showed tissue-specific patterns (Supplemental Figure 4).

With T2T-NIP, we achieved a complete sequence of the important rice reference genome with 385.7 million base pairs (Mbp), including abundant improvements compared with the prior assembly (Figure 1A and Supplemental Tables 4–6). Compared to IRGSP-1.0, T2T-NIP contains 12.5 Mbp of newly identified sequence, including rDNA arrays (33.2%), pericentromeric and centromeric regions (32.1%), transposable elements (27.1%), and telomere and subtelomeric regions (5.1%), all of which are necessary for fundamental cellular processes (Figure 1C–1E). Some of the largest gap-filling regions covered the centromeres of nine chromosomes, subtelomeric and telomeric regions of two chromosomes, and large complex and repetitive regions in three chromosomes, which are represented in IRGSP-1.0 as unknown or unresolved sequences (Figure 1A and Supplemental Table 7). In addition to these apparent gaps, other minor gap regions of IRGSP-1.0 were found to be artificial or otherwise incorrect (Supplemental Table 8). We investigated all possible 500 kb flanking regions adjacent to the 72 major gaps in IRGSP-1.0 and found that most regions far from centromeres and telomeres (39/44) showed excellent synteny with T2T-NIP, while almost all regions close to centromeric gaps (11/12) contained additional minor gaps with extensive large structural differences (e.g., deletions and inversions with lengths >20 kb) compared to T2T-NIP

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

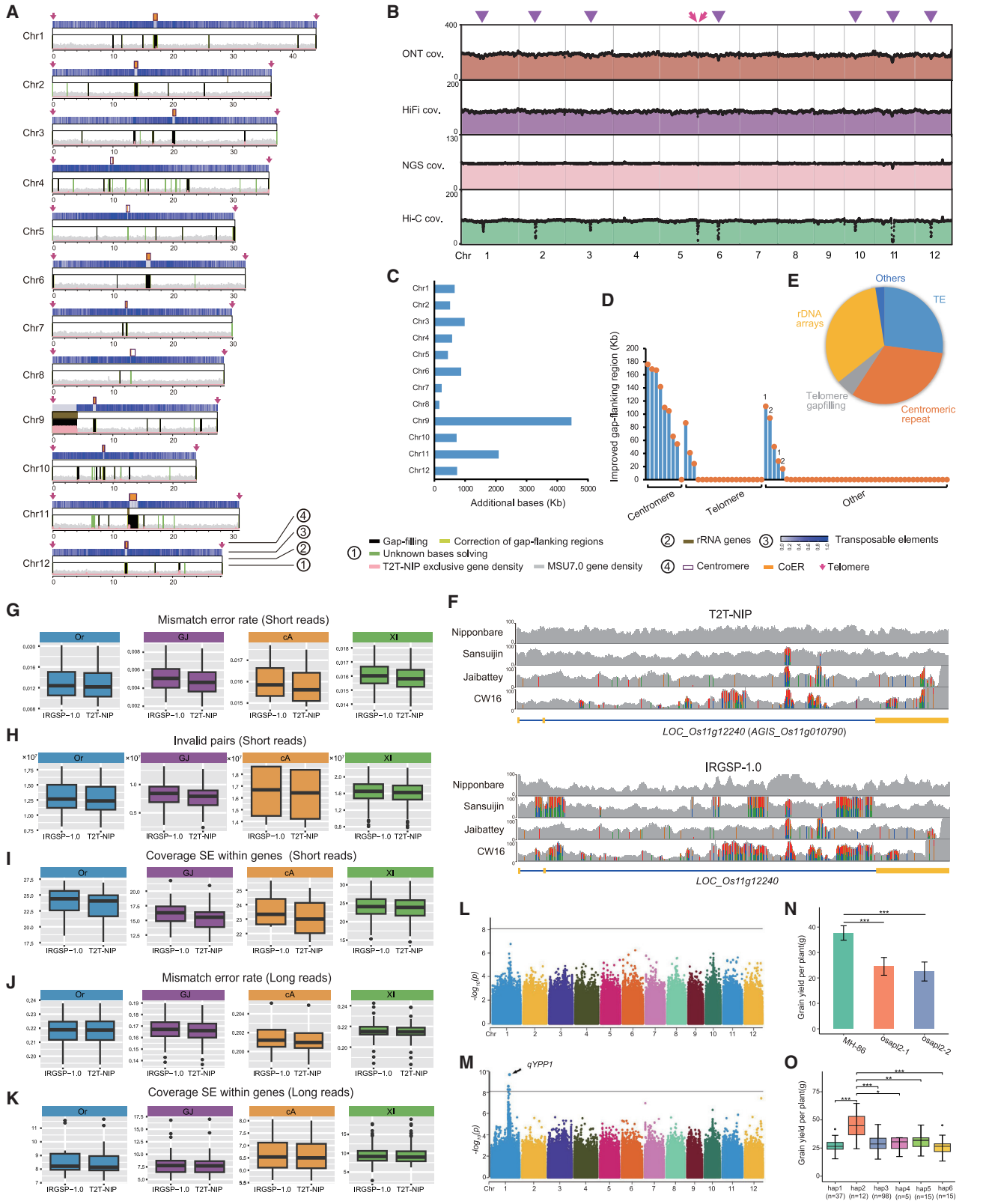


Figure 1. Improvements of the complete rice reference genome T2T-NIP on assembly statistics and potential applications.

(A) Ideogram of T2T-NIP assembly features. For each chromosome (Chr), the following information is provided from bottom to top: gaps and issues in IRGSP-1.0 fixed by T2T-NIP overlaid with the density of genes exclusive to T2T-NIP in red; rRNA genes; transposable elements; denotation of centromeres and telomeres. The bottom scale is measured in Mbp. CoER, CentO-enriched region,

(legend continued on next page)

(Figure 1D). Additionally, four major gaps and their flanking regions with several minor gaps could be well resolved by T2T-NIP, resulting in two continuous regions of 100–117 kb (Figure 1D and Supplemental Table 7). These results demonstrated a significant update of the rice reference genome by resolving the gaps and misassembled structures probably caused by complex and large repetitive structures in IRGSP-1.0.

T2T-NIP removes a long-standing barrier that has hidden 3% of the genome from sequence-based analysis, resolving all centromeric and telomeric regions. Therefore, it is important to further describe the initial analysis of a truly complete rice reference genome and to discuss its potential applications. We have produced a rich collection of annotations and omics datasets for T2T-NIP, including gene models and transposon elements (TEs), RNA sequencing, and methylation datasets, as presented in an online database (http://www.ricesuperpir.com/web/nip).

To highlight the utility of these genetic resources, we demonstrate examples of complex duplicated regions in chromosomes 10 and 11 that were associated with previously unresolved gaps. The gene *AGIS_Os10g035850* (denoted as *LOC_Os10g43075* in IRGSP-1.0/MSU7) traversed across the boundary of a major gap at the subtelomeric region of chromosome 10, resulting in an incomplete annotation of only 76.3% of the entire gene and some misannotated exons in the previous version. T2T-NIP thus supported the correction of this gene model, including an addition of six new exons into each of its two splicing alternatives from the gap-filling region (Supplemental Figure 5). Most TE-related genes have multiple copies (paralogs) caused by repetitive sequences, which previously have always complicated their genetic analysis. When mapping NGS reads, the absence of the additional paralogs in IRGSP-1.0 causes these reads to incorrectly align to *LOC_Os11g12240* (*AGIS_Os11g010790*), resulting in many false-positive variants (Figure 1F). When mapped to T2T-NIP, the reads show the expected coverage and a typical heterozygous variation pattern at a small region. Any variants within these paralogs, and others like them, will be overlooked when using IRGSP-1.0 as a reference, thereby promoting the importance of the release of T2T-NIP.

To investigate how the T2T-NIP affects short-read variant calling, we collected NGS reads of 230 cultivated (*Oryza sativa*) and wild (*Oryza rufipogon*) rice accessions from our previous study (Shang et al., 2022). The cultivated collection consisted of three populations: *Xian/indica* (XI), *Geng/japonica* (GJ), and *Aus* (cA). The same pipeline was applied for variant calling based on T2T-NIP and IRGSP-1.0 to eliminate the interferences caused by software parameters. On average, BWA-MEM mapped an additional 1.04×10^7 (6.9%) of properly paired reads to T2T-NIP compared to IRGSP-1.0. Interestingly, even though more reads align to T2T-NIP, the subsequent per-read mismatch rate was 1.2%–8.2% lower across all populations (Figure 1G). Similarly, T2T-NIP improved other mapping characteristics such as reducing the number of misoriented read pairs (Figure 1H) and improving coverage uniformity (Figure 1I) compared to IRGSP-1.0. Within gene regions, we noted a decrease of 2.0%–4.3% in the standard deviation of read coverage with analogous improvements among all population groups (Figure 1I). From these alignments, we identified a total of 741 895 221 high-quality single-nucleotide variants and small indel variants relative to T2T-NIP (per-sample mean, 3 225 631) compared to 744 667 800 variants relative to IRGSP-1.0 (per-sample mean, 3 237 686), observing a shared decrease in the number of called variants per individual genome (Supplemental Figure 6 and Supplemental Table 9). Along with the improvement in the per-read mismatch rate, we attribute the reduction in the number of per-sample variant calls to the lower number of consensus errors, structural errors, and especially the resolution of the complex repetitive regions with correct copies in T2T-NIP (Figure 1F). This conclusion is supported by the observation that the number of heterozygous variants per sample decreased largely in all populations while their homozygous variants showed a slight increase except for GJ (Supplemental Figure 6 and Supplemental Table 9). These results demonstrated the superiority of T2T-NIP as a reference genome for more accurate mapping and variation analysis based on short reads.

Next, we investigated the effects of using T2T-NIP as a reference genome for structural variant (SV) calling from published long reads (Shang et al., 2022). Alignment to T2T-NIP also reduced the observed mismatch rate per mapped read (Figure 1J) and the standard deviation of coverage within genes (Figure 1K) across all populations. T2T-NIP also corrected structural errors in IRGSP-1.0 and contained a complete assembly of the genome,

(B) Sequencing coverage and assembly validation of the T2T-NIP. Uniform whole-genome coverage of mapped HiFi, ONT, Hi-C, and Illumina reads is shown with primary alignments. Large centromeres caused low coverage of mapped Hi-C reads due to insufficient paired reads from this region but were recovered by other types of sequencing datasets.
(C–E) Additional bases in the T2T-NIP assembly relative to IRGSP-1.0 per chromosome, which was evaluated by total length difference **(C)**, corrected gap-flanking regions **(D)**, and sequence type **(E)**. The digits above the bars denote gap-flanking regions of four gaps that were corrected into two continuous regions.
(F) Reference (gray) and variant (colored) allele coverage for four rice accessions with Illumina reads mapped to the paralog *LOC_Os11g12240*. When mapped to T2T-NIP, the reads show the expected coverage and a typical heterozygous variation pattern at a small region. When mapped to IRGSP-1.0, the region shows excessive coverage and variants, indicating that reads from the missing paralogs are mismapped to the target gene.
(G–K) Improvements to short-read **(G–I)** and long-read **(J** and **K)** mapping and variant calling. Summary of alignment characteristics aligned to T2T-NIP instead of IRGSP-1.0 presented with box plots with the number of invalid pairs, mismatch error rate, and coverage standard error within genes in each accession across different populations. Or, wild rice; XI, *Xian/indica*; GJ, *Geng/japonica*; cA, *Aus*.
(L–O) GWAS analysis of yield per plant based on T2T-NIP as a reference. Manhattan plot shows a significant signal related to a candidate gene *AGIS_Os01g037910* (*OsAGPL2*) based on T2T-NIP **(M)** compared to that based on IRGSP-1.0 **(L)**. Significant differences were observed between yield per plant of different individuals with wild type (MH-86) and function-loss mutation (osapl2-1 and osapl2-2) of *OsAGPL2* **(N)** and between different accessions with different haplotypes of this gene **(O)**. The boxplots in (**G**–**K**) and (**O**) represent the median and the first and third quartiles, and the whiskers extend to minimum and maximum value. Data in (**N**) are presented as mean ± SD of three biological replicates. The asterisks in (**N**) and (**O**) indicate statistical differences according to Student's t test (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

which facilitated a much more accurate alignment, similar to what we observed for short reads (Supplemental Table S10). From these results, we observed a shared reduction (from −16.3% to −4.6%) in the number of SVs from different populations when calling variants against T2T-NIP instead of IRGSP-1.0. Similar to the results of the small variations above, the number of heterozygous variants decreased more than those of homozygous variants (Supplemental Figure 7), likely also due to improvements in resolution of the complex repetitive regions in T2T-NIP, which reduced the rare structures found in IRGSP-1.0.

To supplement our variant and phenotype datasets, we conducted genome-wide association studies (GWASs) to assess potential improvements on efficiency of genetic analysis by using T2T-NIP as a reference genome instead of IRGSP-1.0. A total of 101 associated SNPs were identified for five agronomic traits, in which all associated SNPs were detected only from variant datasets relative to T2T-NIP. For example, a pleiotropic locus related to yield per plant in chromosome 1 (*qYPP1*) of T2T-NIP was significantly associated with both grain yield and plant height that was not identified using IRGSP-1.0 (Figure 1L–1M and Supplemental Figure 8). Gene-editing experiments and phenotype screening revealed significant differences of yield per plant and plant height between plants with wild type and function-loss mutation of a gene encoding the large subunit of ADP-glucose pyrophosphorylase, *OsAGPL2* (Figure 1N and Supplemental Figure 8). A favorable haplotype of *OsAGPL2* was identified, showing significantly higher yield per plant (44.7 ± 11.8 g) than the other haplotypes (Figure 1O). Additionally, we identified some T2T-NIP-specific associated SVs related to grain width (Supplemental Figure 9). These results demonstrated the enhanced efficiency of genetic analysis on population variation and gene mining based on T2T-NIP.

In summary, we achieved complete sequences of the most commonly used rice reference genome in our assembly, T2T-NIP, by addressing the missing 3% of the genomic information, which represents a significant update to this important resource. T2T-NIP introduced ∼12.5 Mbp containing 1324 gene predictions, which include rDNA arrays, centromeric satellite arrays, subtelomeres, and large repeat regions, thereby unlocking these complex regions of the genome for rice variational and functional studies.

DATA AVAILABILITY

All the raw sequencing reads, genome assembly, and annotations for T2T-NIP were deposited in the National Center for Biotechnology Information database under project accession number PRJNA953663 and the National Genomics Data Center database under project accession number PRJCA018610. The genome browser of T2T-NIP and its related annotations and omics datasets can also be easily accessed from our online database website (http://www.ricesuperpir.com/web/nip).

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

This research was supported by the National Natural Science Foundation of China (32188102, 32101718), Guangdong Basic and Applied Basic Research Foundation (2023B1515020053), the Youth Innovation of Chinese Academy of Agricultural Sciences (Y20230C36), and the specific research fund of The Innovation Platform for Academicians of Hainan Province (YSPTZX202303).

AUTHOR CONTRIBUTIONS

Q.Q., W.P. and L.S. conceived the study. W.H., T.W., Y.Y., Q.X., X. Zhao, S.C., X.Li, L.Y. and X.Y. performed data statistical analysis and all bioinformatic analyses. H.Z., X.L., Y.L, W.C., X.W., Bin Z., X.Liu, H.H., H.W., Y.L., C.S., M.G., Z.Z., Bintao Z., Q.Y., H.Q., X.C., Y.C., Q.Z., X.D., C.L., L.G. conducted the field and laboratory experiments for this study. L.S., W.H., T.W., Q.Y., X.Z., and Y.Y. wrote the manuscript with guidance from Q.Q., W.P. and L.S. Y.Z., X. Zheng, J.R. and Z.C. revised the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Weiwei Jin (China Agricultural University) and Dr. Wei Huang (China Agricultural University) for their kind help in conducting the ChIP-seq experiments. The authors have no conflicts to declare.

Received: May 19, 2023
Revised: July 30, 2023
Accepted: August 6, 2023
Published: August 7, 2023

Lianguang Shang[1,2,6,], Wenchuang He[1,6], Tianyi Wang[1,6], Yingxue Yang[1,6], Qiang Xu[1,6], Xianjia Zhao[1,6], Longbo Yang[1], Hong Zhang[1], Xiaoxia Li[1], Yang Lv[1], Wu Chen[1], Shuo Cao[1], Xianmeng Wang[1], Bin Zhang[1], Xiangpei Liu[1], Xiaoman Yu[1], Huiying He[1], Hua Wei[1], Yue Leng[1], Chuanlin Shi[1], Mingliang Guo[1], Zhipeng Zhang[1], Bintao Zhang[1], Qiaoling Yuan[1], Hongge Qian[1], Xinglan Cao[1], Yan Cui[1], Qianqian Zhang[1], Xiaofan Dai[1], Congcong Liu[1], Longbiao Guo[3], Yongfeng Zhou[1], Xiaoming Zheng[4], Jue Ruan[1], Zhukuan Cheng[5], Weihua Pan[1,*] and Qian Qian[1,2,3,*]*

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
[2]Yazhouwan National Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City, Hainan Province 572024, China
[3]State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China
[4]National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[5]Key Laboratory of Plant Functional Genomics of the Ministry of Education, Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou 225009, China
[6]These authors contributed equally to this article.
***Correspondence: Lianguang Shang (shanglianguang@caas.cn), Weihua Pan (panweihua@caas.cn), Qian Qian (qianqian188@hotmail.com)**
https://doi.org/10.1016/j.molp.2023.08.003

REFERENCES

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific

retrotransposon. Plant Cell **14**:1691–1704. https://doi.org/10.1105/tpc.003079.

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature **436**:793–800. https://doi.org/10.1038/nature03895.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice **6**:4. https://doi.org/10.1186/1939-8433-6-4.

Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L.Y., Gao, L.Z., Li, P., and Lu, S. (2021). Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol. Plant **14**:1745–1756. https://doi.org/10.1016/j.molp.2021.06.017.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. **54**:e6. https://doi.org/10.1093/pcp/pcs183.

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., et al. (2022). A super pan-genomic landscape of rice. Cell Res. **32**:878–896. https://doi.org/10.1038/s41422-022-00685-z.

Song, J.M., Xie, W.Z., Wang, S., Guo, Y.X., Koo, D.H., Kudrna, D., Gong, C., Huang, Y., Feng, J.W., Zhang, W., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant **14**:1757–1767. https://doi.org/10.1016/j.molp.2021.06.018.

Zhang, Y., Fu, J., Wang, K., Han, X., Yan, T., Su, Y., Li, Y., Lin, Z., Qin, P., Fu, C., et al. (2022). The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol. J. **20**:1642–1644. https://doi.org/10.1111/pbi.13880.

Molecular Plant, Volume *16*

Supplemental information

A complete assembly of the rice Nipponbare reference genome

Lianguang Shang, Wenchuang He, Tianyi Wang, Yingxue Yang, Qiang Xu, Xianjia Zhao, Longbo Yang, Hong Zhang, Xiaoxia Li, Yang Lv, Wu Chen, Shuo Cao, Xianmeng Wang, Bin Zhang, Xiangpei Liu, Xiaoman Yu, Huiying He, Hua Wei, Yue Leng, Chuanlin Shi, Mingliang Guo, Zhipeng Zhang, Bintao Zhang, Qiaoling Yuan, Hongge Qian, Xinglan Cao, Yan Cui, Qianqian Zhang, Xiaofan Dai, Congcong Liu, Longbiao Guo, Yongfeng Zhou, Xiaoming Zheng, Jue Ruan, Zhukuan Cheng, Weihua Pan, and Qian Qian

Supplemental Information

A complete assembly of the rice Nipponbare reference genome

Lianguang Shang[1,2,#,*], Wenchuang He[1#], Tianyi Wang[1#], Yingxue Yang[1#], Qiang Xu[1#], Xianjia Zhao[1#], Longbo Yang[1], Hong Zhang[1], Xiaoxia Li[1], Yang Lv[1], Wu Chen[1], Shuo Cao[1], Xianmeng Wang[1], Bin Zhang[1], Xiangpei Liu[1], Xiaoman Yu[1], Huiying He[1], Hua Wei[1], Yue Leng[1], Chuanlin Shi[1], Mingliang Guo[1], Zhipeng Zhang[1], Bintao Zhang[1], Qiaoling Yuan[1], Hongge Qian[1], Xinglan Cao[1], Yan Cui[1], Qianqian Zhang[1], Xiaofan Dai[1], Congcong Liu[1], Longbiao Guo[3], Yongfeng Zhou[1], Xiaoming Zheng[4], Jue Ruan[1], Zhukuan Cheng[5], Weihua Pan[1,*], Qian Qian[1,2,3,*]

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China

[2]Yazhouwan National Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City, Hainan Province，572024, China

[3]State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou, 310006, China

[4]National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

[5]Key Laboratory of Plant Functional Genomics of the Ministry of Education, Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou 225009, China.

[#]These authors contributed equally to this work.

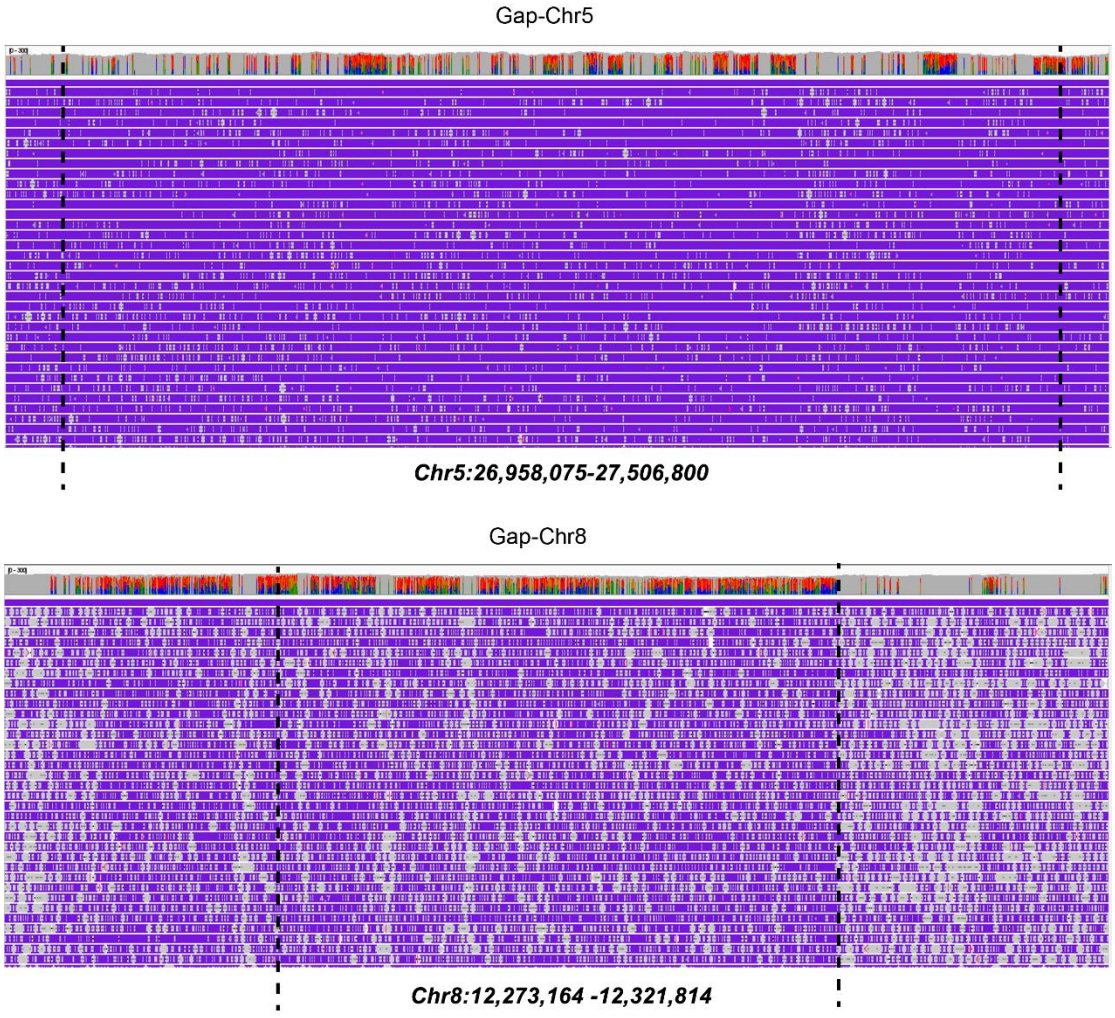
*Correspondence: shanglianguang@caas.cn, panweihua@caas.cn, qianqian188@hotmail.com

This supplemental information file contains:

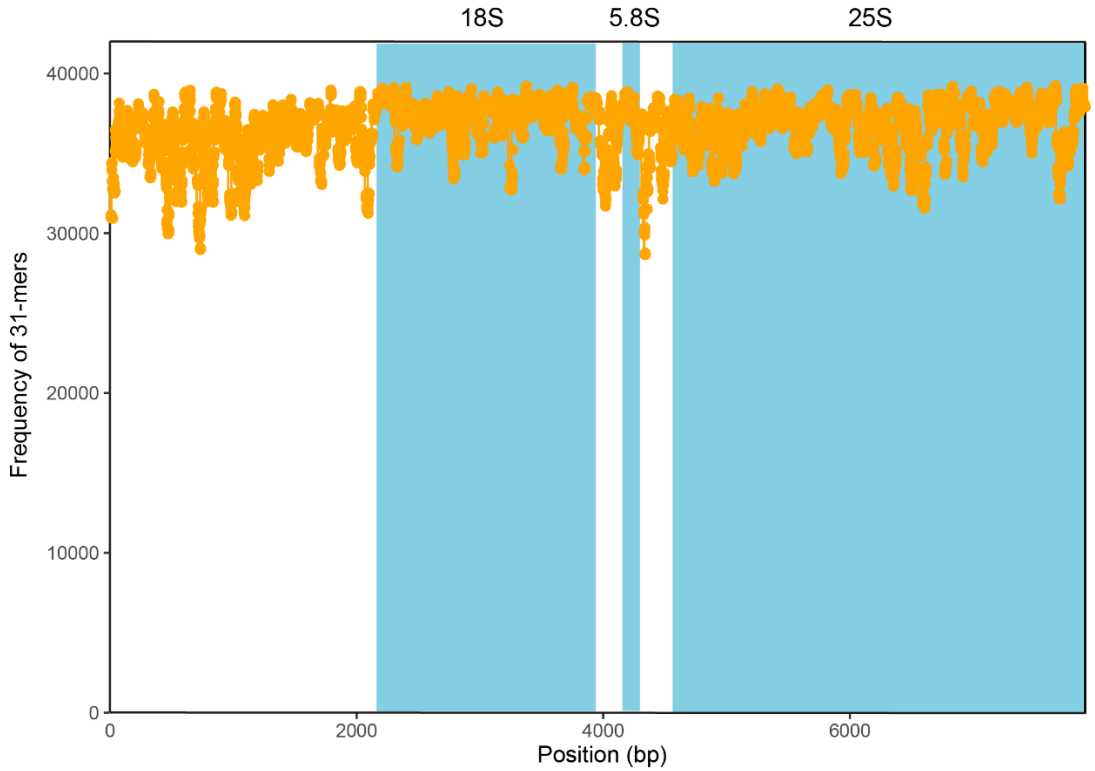
Supplemental figures S1-S9

Supplemental materials and methods

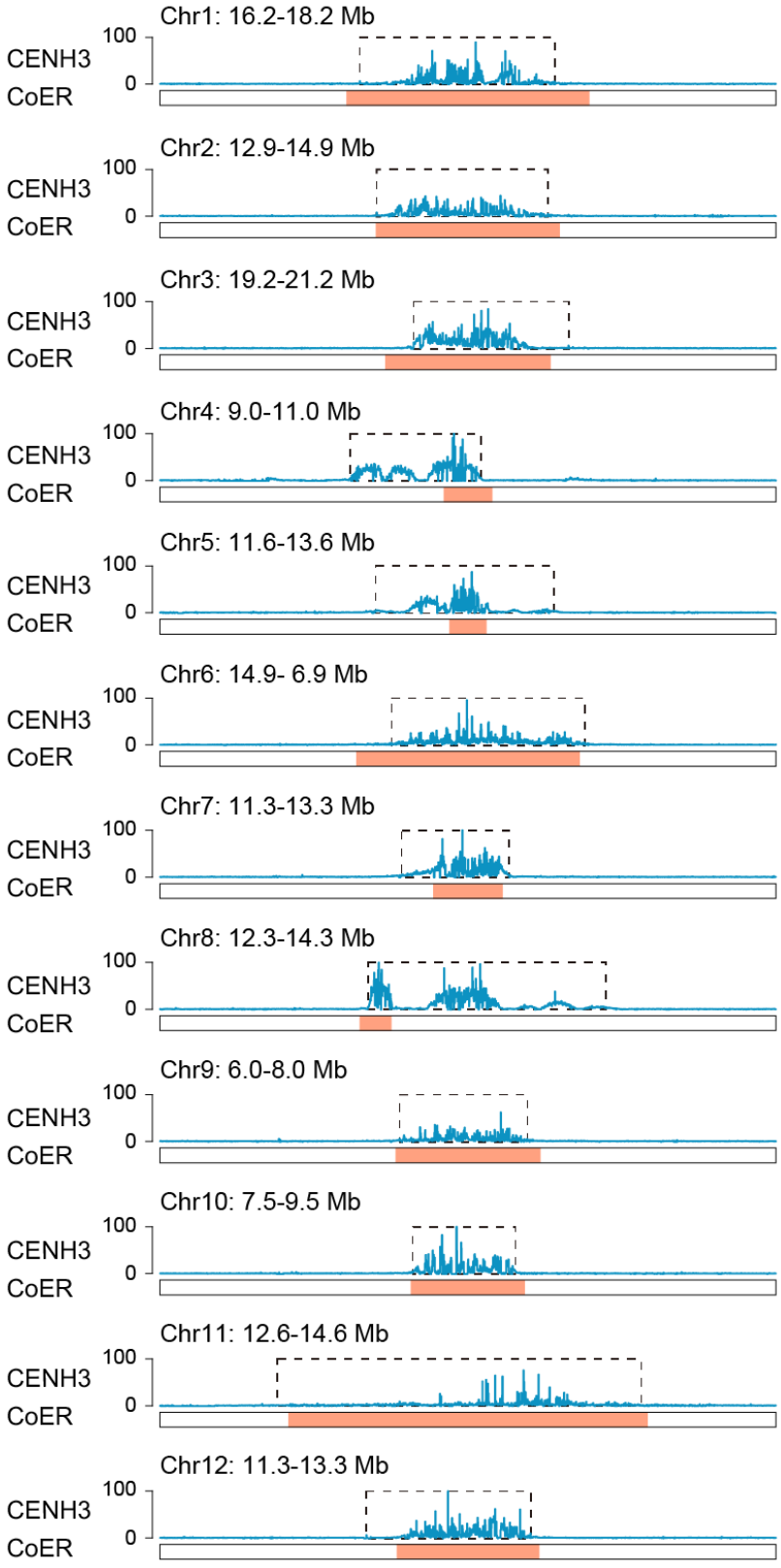
Supplemental figures



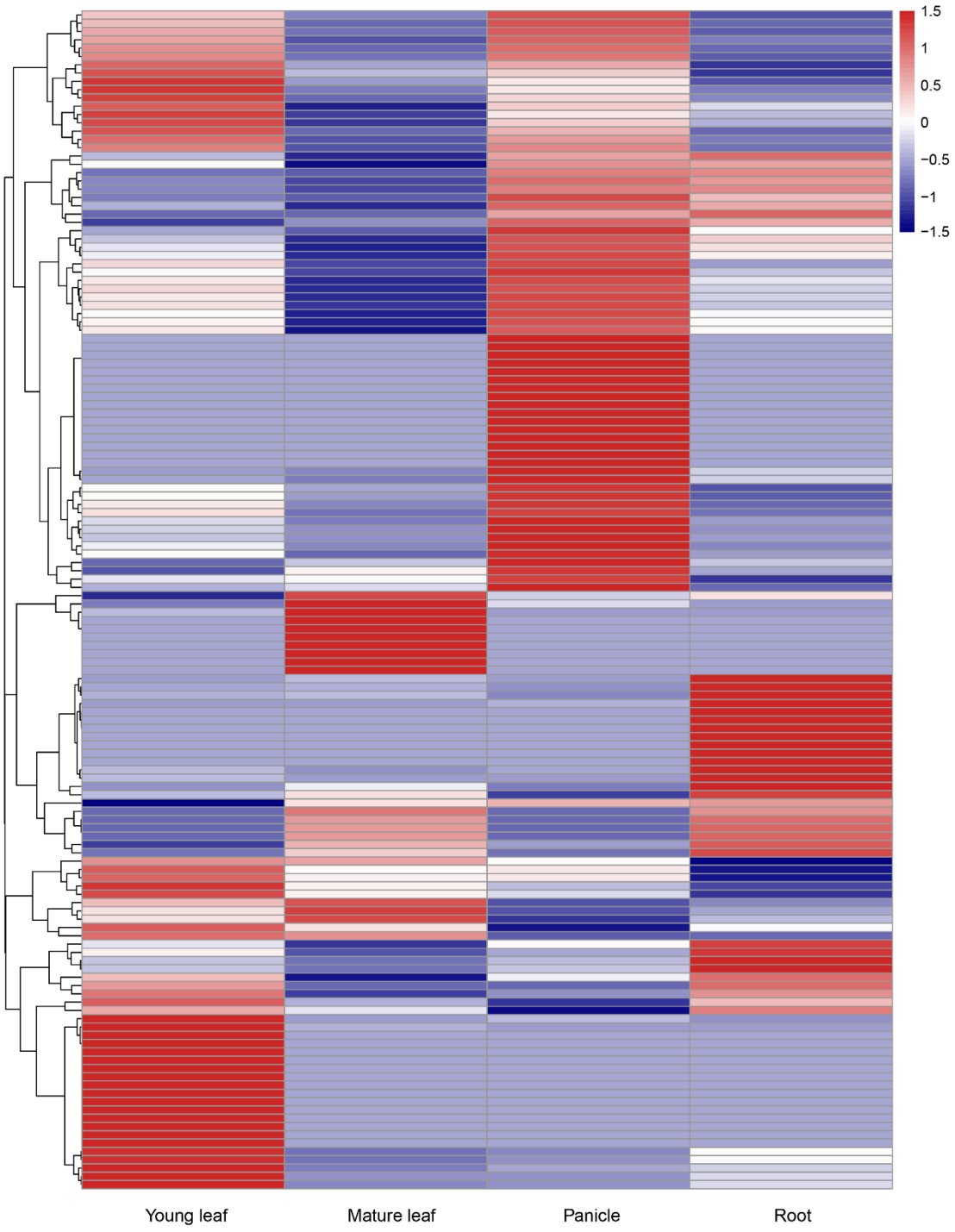
Supplemental figure 1. Nanopore long reads coverage of two closed gap regions in the T2T-NIP. Detailed position of the gap-filling region was noted bellow the coverage plot.



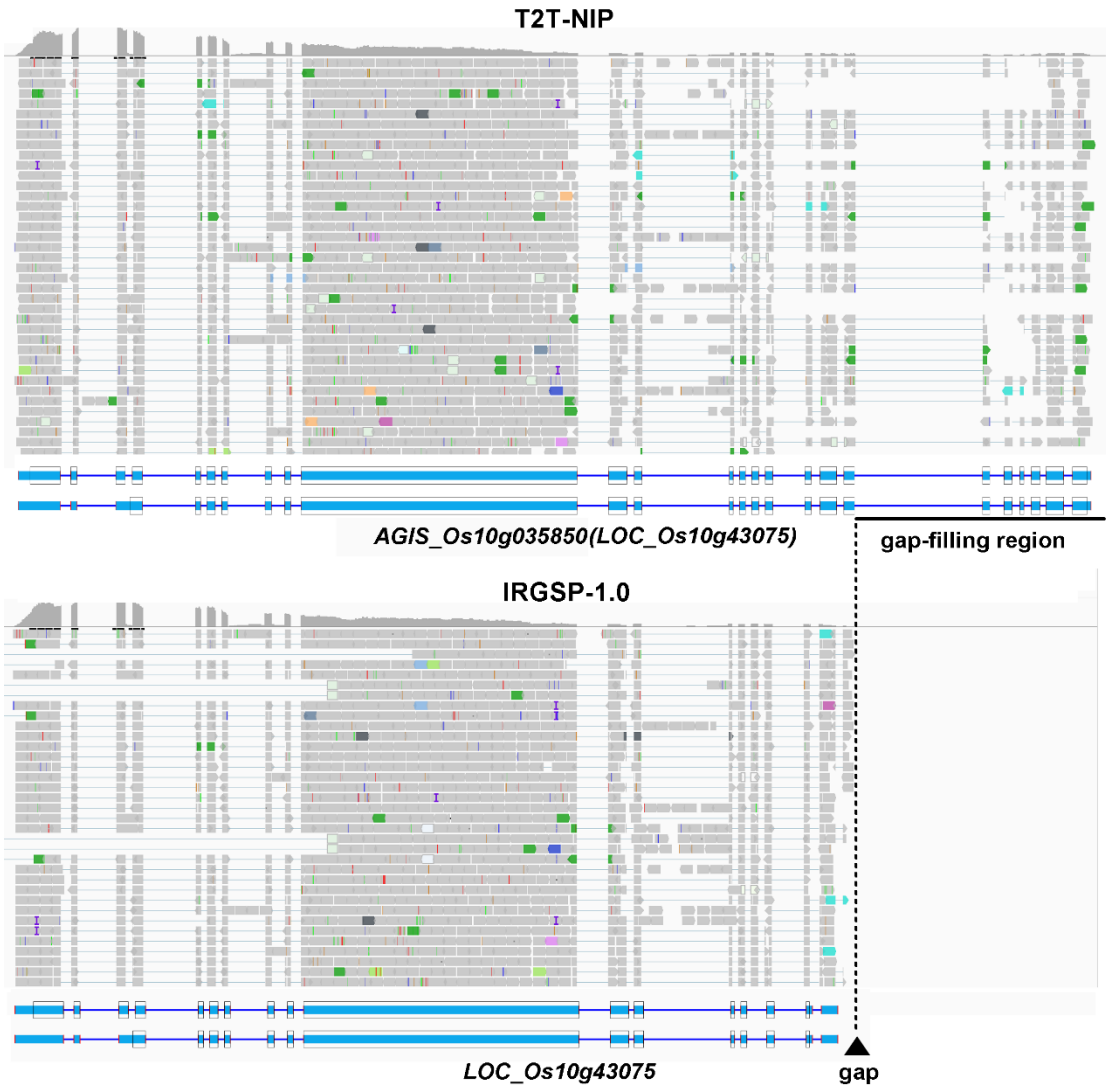
Supplemental figure 2. Frequency of 31-mers across the monomer sequence of 45S rDNA array based on HiFi reads. A relatively low frequency corresponds to small variants or sequencing error existed in the corresponding mer. This plot confirms the strong conservation among the repeats of this 45S rDNA array with only rare variations.



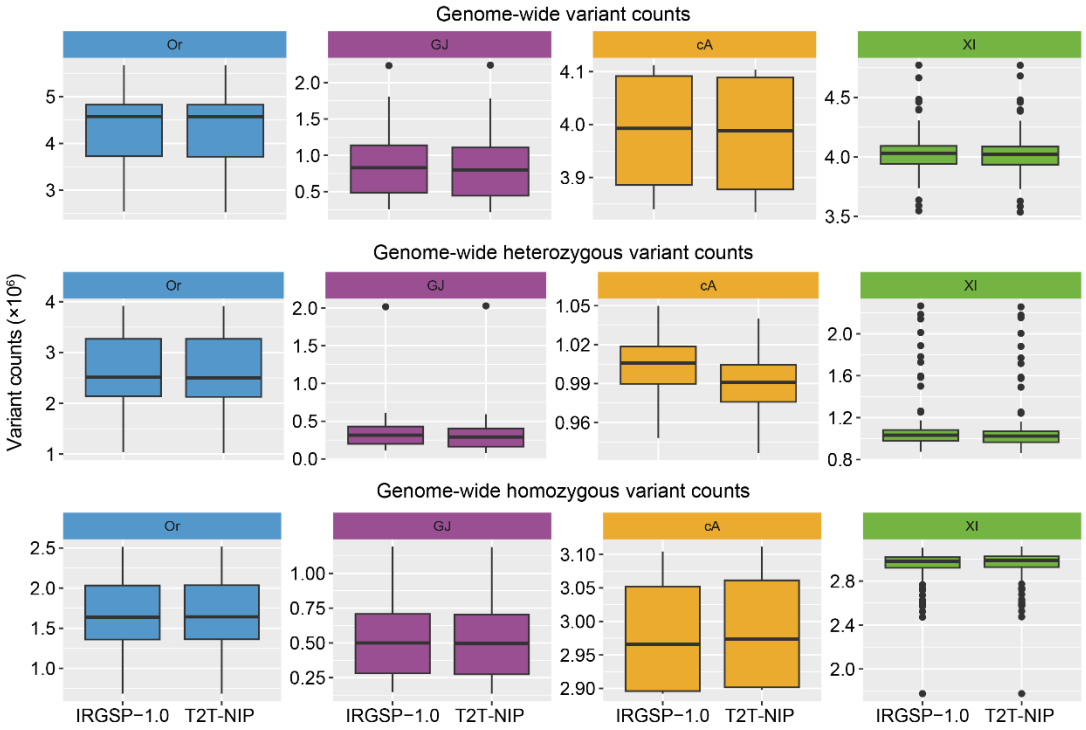
Supplemental figure 3. Schematic representation showing the distribution of different sequence compositions across the 12 centromeres. The CENH3 levels were represented by the enrichment level in 1 kb windows along chromosomes. The centromeres were marked by dotted boxes. CentO-enriched regions (CoERs) were presented bellow each plotting for CENH3 levels.



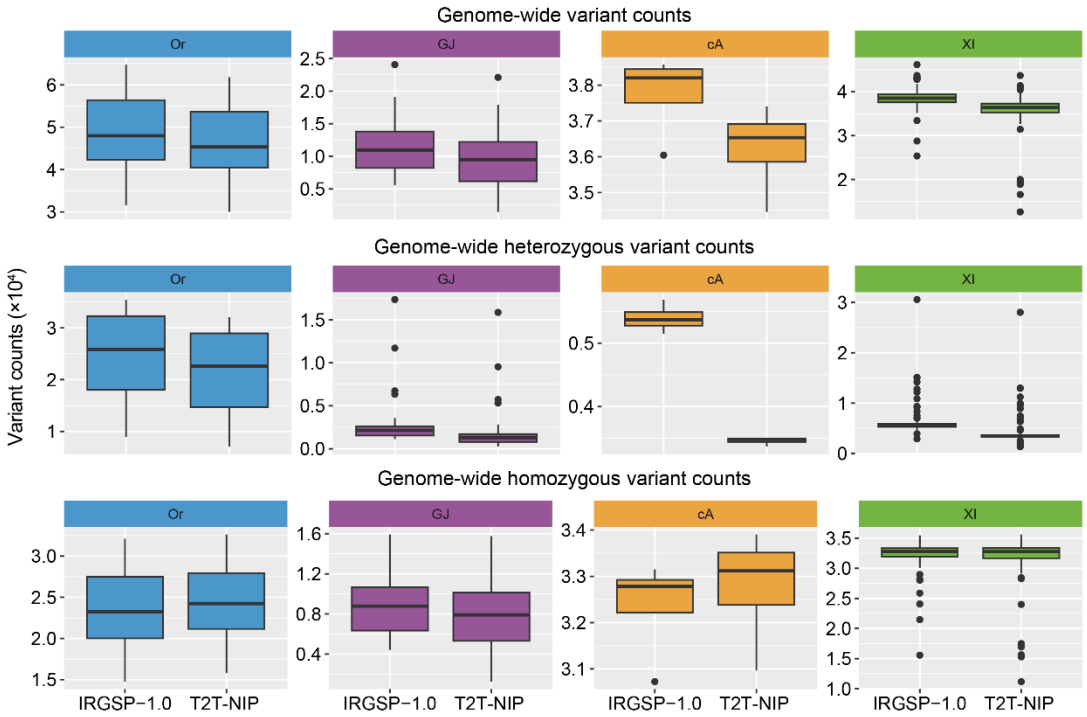
Supplemental figure 4. Expression profile of the newly identified genes in gap-filling regions of T2T-NIP based on transcriptome datasets from different rice tissues. The FPKM values derived from different issues were standardized for each gene as presented in the heat map.



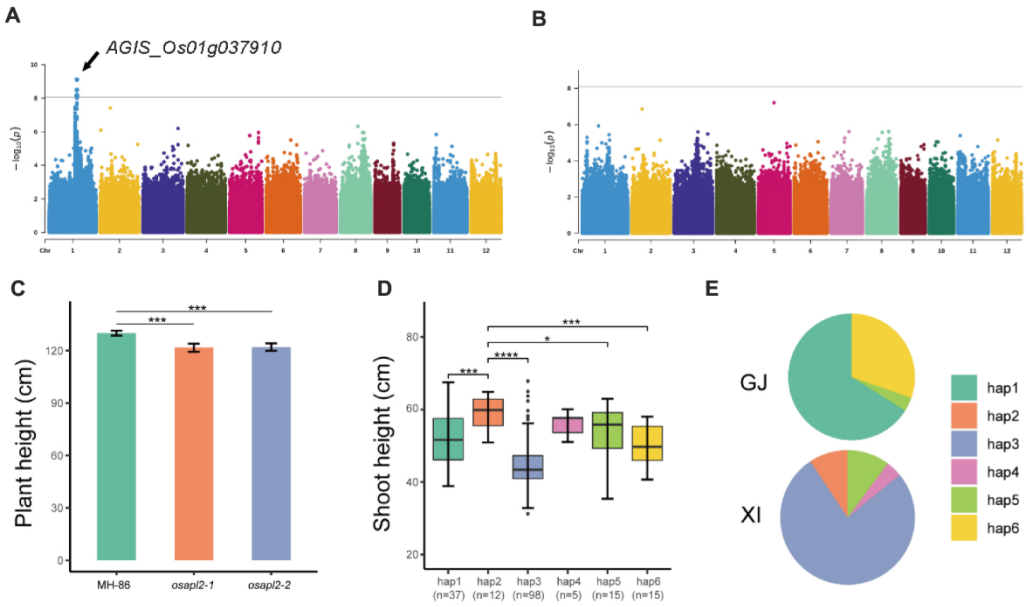
Supplemental figure 5. Correction for a gene model across a gap-filling region. The gene *LOC_Os10g43075* in IRGSP-1.0 was annotated with incomplete gene structure due to a nearby major gap (down), and was corrected in T2T-NIP as *AGIS_Os10g035940* with full annotation by resolving that gap (up). Coverage and alignment of the RNA-seq reads for the two gene annotations were shown above the corresponding gene models.



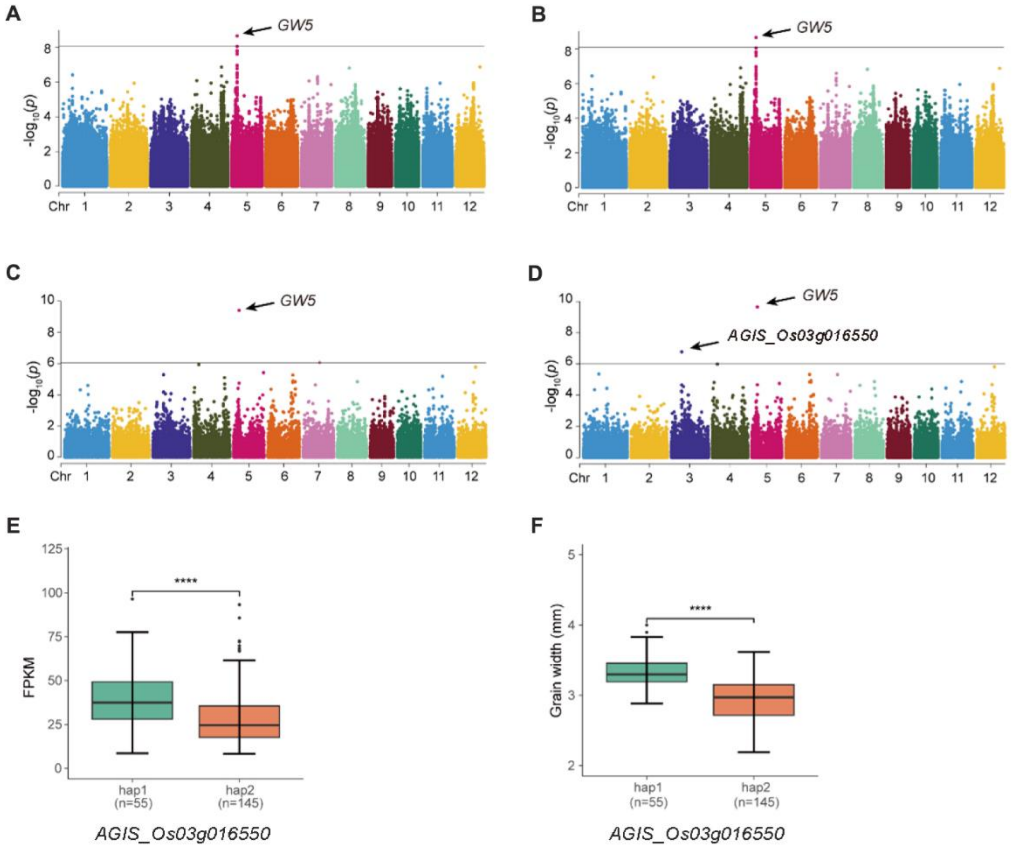
Supplemental figure 6. Boxplots of overall number of small variants, number of heterozygous and homozygous variants found in each rice accession across different populations based on Illumina reads. Comparison of different results with IRGSP-1.0 and T2T-NIP as a reference genome was presented. Or: wild rice; XI: *Xian*/*indica*, GJ: *Geng*/*japonica*, cA: *Aus*. The boxplots represent the median and the first and third quartiles, and the whiskers extend to minimum and maximum value.



Supplemental figure 7. Boxplots of overall number of structural variants, number of heterozygous and homozygous variants found in each rice accession across different populations based on Nanopore ONT reads. Comparison of different results with IRGSP-1.0 and T2T-NIP as a reference genome was presented. Or: wild rice; XI: *Xian*/*indica*, GJ: *Geng*/*japonica*, cA: *Aus*. The boxplots represent the median and the first and third quartiles, and the whiskers extend to minimum and maximum value.



Supplemental figure 8. Functional diversity of the candidate gene *AGIS_Os01g037910*. A-B, Manhattan plot showed a significant signal related to a candidate gene *AGIS_Os01g037910* (*OsAGPL2*) based on T2T-NIP (A) compared to that based on IRGSP-1.0 (B). C-D, Morphological traits of different individuals with and without loss-of-function variation of the candidate gene (C) and different accessions with different haplotypes (D). E, Frequencies of different haplotypes of the candidate gene in XI and GJ populations. Data in (C) are presented as mean ± SD of three biological replicates. The boxplots in (D) represent the median and the first and third quartiles, and the whiskers extend to minimum and maximum value. The asterisks in (C) and (D) indicate statistical differences according to Student's T Test (*, P<0.05; **, P<0.01; ***, P < 0.001).



Supplemental figure 9. GWAS analysis on grain width based on different SNV and SV datasets which were generated by using IRGSP-1.0 and T2T-NIP as a reference genome, respectively. A, Manhattan plot of associated loci related to grain width based on SNPs from using IRGSP-1.0 as reference genome. B, Manhattan plot of associated loci related to grain width based on SNPs from using T2T-NIP as reference genome. C, Manhattan plot of associated loci related to grain width based on SVs from using IRGSP-1.0 as reference genome. D, Manhattan plot of associated loci related to grain width based on SVs from using T2T-NIP as reference genome. E, Gene express patten of the candidate gene *AGIS_Os03g016550* (*LOC_Os03g18770*) among different accessions with different haplotypes. F, Grain width performance of different accessions with different haplotypes. The boxplots in (E) and (F) represent the median and the first and third quartiles, and the whiskers extend to minimum and maximum value. The asterisks indicate statistical differences according to Student's T Test (*, P<0.05; **, P<0.01; ***, P < 0.001).

Supplemental materials and methods

Sample collection and genome sequencing

Young leaves, stems, panicles and roots of Nipponbare were flash-frozen in liquid nitrogen. The leaves were prepared for DNA sequencing and all the four tissues were prepared for RNA sequencing. Genomic DNA and RNA were isolated using the DNeasy Plant Mini kit (Qiagen) following the manufacturer's instructions. For PacBio HiFi sequencing, two single-molecule real-time cells were sequenced on a PacBio Sequel II platform, and a total of 33.0 Gb of HiFi read was generated using CCS (https://github.com/PacificBiosciences/ccs) with the default parameter for the sequenced accessions. For Nanopore ONT sequencing, two cells of DNAs were sequenced on Oxford Nanopore Technology GridION X5/PromethION sequencer and generated a total of 85.6 Gb ONT reads. Illumina Novaseq sequencing platform was used for generating a total of 36.4 Gb Hi-C read dataset and 24.6 Gb paired-end read dataset. For RNA sequencing, Illumina NovaSeq 6000 platform was used to generate a total of 11.2 Gb paired-end reads, and Oxford Nanopore PromethION sequencer was used to generate a total of 9.9 Gb ONT reads.

T2T genome assembly

Both PacBio HiFi reads and Nanopore ONT sequencing reads were self-corrected, trimmed and assembled by HiFiasm v0.18.5-r500 (Cheng et al., 2021; Cheng et al., 2022) to generate contigs (draft assembly v1.0) with high quality and long continuity. The Hi-C sequencing data were applied to hierarchically cluster the assembled contigs into 12 pseudo-chromosomes by using 3d-DNA (Dudchenko et al., 2017). The obtained assembly (v1.1) had 7 gaps and an absence of one telomere across all the chromosomes. To filling the unsolved region, we applied customized strategy to process gaps in complex regions. The published gap-free genome, ZS97RS3 was used to guide the gap-filling process. The SyRI v1.6.4 (Goel et al., 2019) was applied to analyze the synteny between the draft assembly v1.1 and the reference genome to find the corresponding region of reference genome related to the gaps, which were further analyzed in RAfilter (Yang et al., 2023) to calculate gap-region specific kmer sets (k=21) and enrich the target reads containing those specific kmers. Those target reads were re-assembled into one or more contigs covering across the gap with HiFiasm to fill the target gaps. Two gaps were successfully filled with this procedure and generated the assembly version v1.2. For the remaining 5 gaps containing complex repeats, we extended the target region (gap-filling region) to the blanking regions of the gaps and subsequently collected the target reads with the target-region specific kmers as described above. Those obtained reads were also re-assembled in HiFiasm to generate target contigs for filling the gaps. All the remaining 5 gaps were successfully solved with this procedure and produced the assembly version v1.3. For the only absent telomere in Chromosome 9 (Chr9), we firstly conducted a simply extension from the terminal of Chr9 to the corresponding telomere by identifying and mapping HiFi reads with telomere-specific repeats to this region of the assembly v1.3. Verification analysis of this extension was conducted by checking the read coverage of ONT reads with minimap2 v2.21 (Li, 2018) and IGV browser (https://github.com/igvteam/igv), which revealed a much higher coverage depth of ONT reads in this subtelomeric region than the average depth of the whole genome, indicating that this region could be an unsolved large repetitive sequence. Sequence analysis for this high-coverage region showed that it was an rDNA array consisting of multiple exact repeats of 45S rDNA and internal transcribed spacer (ITS) region. We

further mapped 5S, 5.8S and 25S rDNA sequences of Nipponbare (https://rapdb.dna.affrc.go.jp) to the entire assembly v1.3 with blastn program (Altschul et al., 1990) to screen all potential 45S rDNA array and confirmed that this high-coverage region was the only 45S rDNA array in the genome. We then conducted a second extension from both flanking region of this region based on the gap-closure approach for 45S rDNA array as described in maize (Chen et al., 2023). The extension results further confirmed that this region was a large repetitive sequence consisting of nearly-identical tandem repeats of the 45S rDNA and ITS region with rare variations which could not be extended to a valid terminal with the ONT and HiFi reads. Finally, we summed up the length of the 45S rDNA array harbored for ONT and HiFi reads and normalized it by division by the average genome coverage of the two datasets, which resulted a total of 4.25 and 3.77 Mb for this region, respectively. Based on the average size of the two estimations as well as the adjacent regions between the rDNA array and its flanking regions, the total length of this rDNA array was estimated as 4.05 Mb. Therefore, the model array sequences were created by filling in the gaps between confidently assembled boundary repeat units of 45S rDNA array with the estimated copies as described in assembling complete genome of human (Nurk et al., 2022). After filling all the gaps, we generated a gapless telomere-to-telomere assembly version v1.4. Nextpolish v1.3.0 (Hu et al., 2020) with parameters 'HiFi_options -min_read_len 5k -max_read_len 60k -max_depth 90' was used to polish the assembly v1.4 by PacBio HiFi and Illumina short reads, and generated the final complete genome, T2T-NIP.

Verification for accuracy and completeness of the T2T genome

To estimate the coverage depth of the raw reads to the whole genome, we applied BWA v0.7.17 (Li and Durbin, 2009) for mapping Hi-C and next generation sequencing datasets to the T2T-NIP, and applied minimap2 v2.21 (Li, 2018) for mapping PacBio HiFi and Nanopore ONT reads to this complete genome. Depth function of samtools v1.13 (Danecek et al., 2021) was applied to calculate the coverage depth of each site across the genome which was further plotted with a sliding window size of 1 Mb and step size of 100 Kb. To evaluate the consensus accuracy of Nipponbare assembly, Illumina short reads were aligned to its assembly with BWA v0.7.17 (Li and Durbin, 2009) with default parameters and then GATK v4.1.9.0 (Van der Auwera et al., 2013) was used to call SNPs. The homozygous SNPs sites were used to calculate the single base error rate. BUSCO v5.4.3 (Simao et al., 2015) was applied to yield raw statistics of mapped genes with multiple strategies, i.e., "metauk+hmmer" and "augustus+blast", and the results were merged to evaluate all mapped genes in the 1614 genes dataset. The remaining absent genes were further searched in T2T-NIP with tblastn program (Altschul *et al.*, 1990) for identify fragmental mapped genes in them. All those results were used to calculate the final BUSCO statistics of T2T-NIP. The public WGBS data of Nipponbare (Dong et al., 2017) was used to obtain methylation sites. WGBS raw reads were trimmed with Trimmomatic v0.36 (Bolger et al., 2014) and then mapped to T2T-NIP with Bowtie2 v2.1.0 (Langmead et al., 2019) and kept only unique mapped reads by using the tool deduplicate_bismark in bismark v0.22.1 (Krueger and Andrews, 2011). After that, we extracted methylated cytosines using bismark_methylation_extractor and only kept sites with more than four mapped reads.

The annotation of genes and TEs

Multiple strategies were applied to generate high-quality gene annotation for T2T-NIP. Liftoff

v1.6.3 (Shumate and Salzberg, 2021) was primarily used for mapping the previous high-quality gene annotation (MSU7) of IRGSP-1.0 to T2T-NIP with an strict standard (coverage = 100% and identity > 90%). Additional gene copies not annotated in the previous reference could also be detected with this software under filtering params of coverage = 100% and identity > 99%. Genes of the gap-related regions of T2T-NIP were further annotated by using the Genome-Wide-Annotation-Pipeline (https://github.com/unavailable-2374/Genome-Wide-Annotation-Pipeline) with support of *de novo* prediction, RNA sequencing dataset and homologous gene datasets. After that, all gene annotations were manually checked to fix incomplete gene models or repeated annotations. Gene models in the gap-related regions were further manually checked and adjusted with support of RNA sequencing dataset in IGV-GSAman (https://gitee.com/CJchen/IGV-sRNA). Coding sequences of all annotated genes were mapped to the TE library rice6.9.5.liban (https://github.com/oushujun/riceTElib) by using blastn program (Altschul *et al.*, 1990) with a filtering params of e < 1e-10 to identify TE-related genes. The Extensive *de novo* TE Annotator v1.9.6 (Ou et al., 2019) was used to predict TEs with the curated TE library (rice 6.9.5.liban) from its package with parameters '-overwrite 1 -sensitive 1 -anno 1'. The rRNA genes were identified by searching the genome against 5S, 5.8S, 18S and 25S rRNA genes of Nipponbare with blastn program (Altschul *et al.*, 1990), and the hits with coverage > 90% and e-value < 1e-10 were annotated as rRNA genes.

Genome comparison between IRGSP-1.0 and T2T-NIP

To compare the genomic sequences of IRGSP-1.0 and T2T-NIP, we aligned the two assemblies by using minimap2 v2.21 (Li, 2018) and index the alignment BAM file by using samtools v1.13 (Danecek *et al.*, 2021). SyRI (Goel *et al.*, 2019) was used to detect structural variations between the two genomes. To identify potential gaps in IRGSP-1.0 comparing to T2T-NIP, we applied customized Perl script to find the gap-filling regions and its blanking regions. Flanking regions of 500 Kb around each of the previously reported 72 major gaps (Sakai et al., 2013) were extracted from IRGSP-1.0 with seqkit v2.4 (Shen et al., 2016) and mapped to T2T-NIP with minimap2 v1.13 (Danecek *et al.*, 2021). The closer breakpoint of a closest alignment with size > 5Kb to the gap was considered as the bounder of the gap-filling region. The closer breakpoint of a closest alignment with size > 100Kb to the gap was considered as the bounder of corrected flanking region of the corresponding gap. The bedtools v2.30 (Quinlan and Hall, 2010) were applied to compare the aligned positions from minimap2 results of the two whole genome assemblies and the entire regions of all chromosomes in T2T-NIP to identify the unaligned regions with length > 1Kb as potential newly solved regions.

The identification of telomeres and centromeres

We applied seqkit v2.4 (Shen *et al.*, 2016) to search for the whole genome with the known telomere repeat units of rice genome, "CCCTAAA", and identified the longest tandem repeat region containing this unit as the telomere. We used BLAST (Altschul *et al.*, 1990) to align CentO satellite repeats to the genome sequence with an E-value threshold of 1e-5. To more precisely map the core centromere regions, we utilized Tandem Repeats Finder (Benson, 1999) to identify consecutive CentO satellite repeat clusters, occurring no less than twice within a continuous window of 50 kb.

To further analyze the location and sequence of functional centromeres, we perfomed ChIP with rice anti-CENH3 antibody which was produced according as previous description (Nagaki et al., 2004). Briefly, leaves of 4-week-old seedlings were collected and frozen in liquid nitrogen. Nuclei isolation was performed according to the published protocol (Zhang et al., 2012). Extracted nuclei were digested with micrococcal nuclease (Sigma, N3755). The digested mixture was then used for ChIP using the anti-CENH3 antibody. The rabbit pre-immune blood serum was used as negative control. ChIP-seq libraries was constructed following standard protocols (Nagaki et al., 2003). The ChIP–seq libraries were sequenced (Paired-end 150 nt) on Illumina platforms in Novogene Bioinformatics Technology Co., Ltd (Beijing, China). Resulting ChIP–seq data of CENH3 was used for the identification of functional centromere regions of the T2T-NIP. Raw reads were subjected to adapter trimming and filtering by fastp v0.23.1 (Chen et al., 2018). Resulting ChIP–seq reads were mapped onto the T2T-NIP using Bowtie2 v2.1.0 (Langmead *et al.*, 2019). Uniquely mapped reads were extracted by samtools v1.13 (Danecek *et al.*, 2021) for further analysis. Enrichment level of CENH3 for each base was obtained using bamCompare in the Deeptools package (v3.5.1). Average enrichment of each 1 kb-bin of the genome was then calculated. The bins with enrichment levels greater than 5 were retained and with a distance interval less than 200Kb were merged. The final centromeric regions were determined by visual inspection of the distribution of CENH3 ChIP–seq peaks.

Population variation calling and GWAS analysis

Illumina and Nanopore sequencing data of 251 cultivated and wild rice accessions were collected from our previous study (Shang et al., 2022). Quality control of short sequencing reads were conducted by using Trimmomatic (10) (version: 0.39 parameter: MINLEN: 75 LEADING: 20 TRAILING: 20 SLIDINGWINDOW: 5:20). The reads were mapped to the two versions of rice reference genome, IRGSP-1.0 and T2T-NIP, with BWA v0.7.17 (Li and Durbin, 2009) and then were used for SNP calling in Sentieon v202112.02 (Kendig et al., 2019). The long sequencing reads were mapped to the two assemblies with minimap2 v1.13 (Danecek *et al.*, 2021) and the low-quality reads (MQ < 20) were filtered with samtools v1.13 (Danecek *et al.*, 2021). The cuteSV v2.0.3 (Jiang et al., 2020) were further used for SV calling.

Vcftools v0.1.16 (Danecek et al., 2011) was used to filter SNPs retaining sites with parameters “-maf 0.05 -max-missing 0.8”. Principal component analysis (PCA) was conducted by using plink v1.90b6.26 (Purcell et al., 2007). The kinship of the population was estimated by GEMMA v0.98.5 (Zhou and Stephens, 2012). GWAS analysis was performed for the cultivated rice population by using GEMMA v0.98.5 (Zhou and Stephens, 2012) with a mixed linear model and the first five principal components and standardized matrix of kinship were used as covariates. The threshold for GWAS was calculated using Bonferroni test (0.05/number of SNPs). Phenotypes in our dataset comprised yield per plant and plant height were surveyed at the harvest stage in Mianyang of Sichuan provience, China (2020) and at heading stage in green house (2019), respectively. For construction of CRISPR/Cas9 plant expression vectors and rice transformation, one fragment (CTTGTGCACAACCAATGAGAAGG) targeting *OsAPL2* was selected to generate mutants using the CRISPR/Cas9 system with the plant expression vector VK005-01 (ViewSolid Biotech, Beijing, China) (Zhu et al., 2019). The recombinant plasmids were inserted into

Agrobacterium tumefaciens strain EHA105 cells for the subsequent transformation of rice varieties MH-86. Plant heights of the MH86 and the transgenic plants were measured at the maturation stage, and yield per plant were investigated at the harvest stage in the field. Expression levels of different candidate genes were evaluated based on transcriptome datasets reported in our previous study (Shang *et al.*, 2022). Haplotypes of the candidate genes were estimated based on the population SNPs which were generated as described above and were then filtered by applying a MAF $\geq$ 0.05 cut-off. Low-frequency haplotypes could arise owing to errors in genotypes calling and we therefore filtered the haplotypes by retaining the ones with frequency $\geq$ 3. T-test was applied to assess the different significance between phenotypes and gene expression levels of different rice accessions and individuals.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215**:403-410. 10.1016/S0022-2836(05)80360-2.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27**:573-580. 10.1093/nar/27.2.573.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114-2120. 10.1093/bioinformatics/btu170.

Chen, J., Wang, Z., Tan, K., Huang, W., Shi, J., Li, T., Hu, J., Wang, K., Wang, C., Xin, B., et al. (2023). A complete telomere-to-telomere assembly of the maize genome. Nat Genet **55**:1221-1231. 10.1038/s41588-023-01419-6.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**:i884-i890. 10.1093/bioinformatics/bty560.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods **18**:170-175. 10.1038/s41592-020-01056-5.

Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.P., Urban, L., Gemmell, N.J., and Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol

40:1332-1335. 10.1038/s41587-022-01261-x.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience **10**10.1093/gigascience/giab008.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics **27**:2156-2158. 10.1093/bioinformatics/btr330.

Dong, P., Tu, X., Chu, P.-Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. Molecular Plant **10**:1497-1509. https://doi.org/10.1016/j.molp.2017.11.005.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science **356**:92-95. 10.1126/science.aal3327.

Goel, M., Sun, H., Jiao, W.B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol **20**:277. 10.1186/s13059-019-1911-0.

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36**:2253-2255. 10.1093/bioinformatics/btz891.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol **21**:189. 10.1186/s13059-020-02107-y.

Kendig, K.I., Baheti, S., Bockol, M.A., Drucker, T.M., Hart, S.N., Heldenbrand, J.R., Hernaez, M., Hudson, M.E., Kalmbach, M.T., Klee, E.W., et al. (2019). Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. Front Genet **10**:736. 10.3389/fgene.2019.00736.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27**:1571-1572. 10.1093/bioinformatics/btr167.

Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics **35**:421-432. 10.1093/bioinformatics/bty648.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**:3094-3100. 10.1093/bioinformatics/bty191.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760. 10.1093/bioinformatics/btp324.

Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., and Jiang, J. (2003). Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. Genetics **163**:1221-1225. 10.1093/genetics/163.3.1221.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R., and Jiang, J. (2004). Sequencing of a rice centromere uncovers active genes. Nat Genet **36**:138-145. 10.1038/ng1289.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human

genome. Science **376**:44-53. 10.1126/science.abj6987.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol **20**:275. 10.1186/s13059-019-1905-y.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**:559-575. 10.1086/519795.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841-842. 10.1093/bioinformatics/btq033.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol **54**:e6. 10.1093/pcp/pcs183.

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., et al. (2022). A super pan-genomic landscape of rice. Cell Res **32**:878-896. 10.1038/s41422-022-00685-z.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One **11**:e0163962. 10.1371/journal.pone.0163962.

Shumate, A., and Salzberg, S.L. (2021). Liftoff: accurate mapping of gene annotations. Bioinformatics **37**:1639-1643. 10.1093/bioinformatics/btaa1016.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015).

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210-3212. 10.1093/bioinformatics/btv351.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics **43**:11 10 11-11 10 33. 10.1002/0471250953.bi1110s43.

Yang, J., Zhao, X., Jiang, H., Yang, Y., Hou, Y., and Pan, W. (2023). RAfilter: an algorithm for detecting and filtering false-positive alignments in repetitive genomic regions. Hortic Res **10**:uhac288. 10.1093/hr/uhac288.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. Genome Res **22**:151-162. 10.1101/gr.131342.111.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet **44**:821-824. 10.1038/ng.2310.

Zhu, Y., Lin, Y., Chen, S., Liu, H., Chen, Z., Fan, M., Hu, T., Mei, F., Chen, J., Chen, L., et al. (2019). CRISPR/Cas9-mediated functional recovery of the recessive rc allele to develop red rice. Plant Biotechnol J **17**:2096-2105. 10.1111/pbi.13125.