Molecular Plant
Research Report

CellPress
Partner Journal

Two gap-free reference genomes and a global view of the centromere architecture in rice

Jia-Ming Song[1,2,8], Wen-Zhao Xie[1,8], Shuo Wang[1,8], Yi-Xiong Guo[1], Dal-Hoe Koo[3], Dave Kudrna[4], Chenbo Gong[1], Yicheng Huang[1], Jia-Wu Feng[1], Wenhui Zhang[1], Yong Zhou[5], Andrea Zuccolo[5], Evan Long[6], Seunghee Lee[4], Jayson Talag[4], Run Zhou[1], Xi-Tong Zhu[1], Daojun Yuan[1], Joshua Udall[6,9], Weibo Xie[1], Rod A. Wing[4,5,7], Qifa Zhang[1], Jesse Poland[3,*], Jianwei Zhang[1,*] and Ling-Ling Chen[1,2,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

[2]College of Life Science and Technology, Guangxi University, Nanning 530004, China

[3]Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

[4]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

[5]Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[6]Plant and Wildlife Science Department, Brigham Young University, Provo, UT 84602, USA

[7]International Rice Research Institute (IRRI), Strategic Innovation, Los Baños, 4031 Laguna, Philippines

[8]These authors contributed equally

[9]Present address: Crop Germplasm Research Unit, USDA-ARS, 2881 F&B Road, College Station, TX 77845, USA

***Correspondence: Jesse Poland (jpoland@ksu.edu), Jianwei Zhang (jzhang@mail.hzau.edu.cn), Ling-Ling Chen (llchen@mail.hzau.edu.cn)**

https://doi.org/10.1016/j.molp.2021.06.018

ABSTRACT

Rice (*Oryza sativa*), a major staple throughout the world and a model system for plant genomics and breeding, was the first crop genome sequenced almost two decades ago. However, reference genomes for all higher organisms to date contain gaps and missing sequences. Here, we report the assembly and analysis of gap-free reference genome sequences for two elite *O. sativa xian/indica* rice varieties, Zhenshan 97 and Minghui 63, which are being used as a model system for studying heterosis and yield. Gap-free reference genomes provide the opportunity for a global view of the structure and function of centromeres. We show that all rice centromeric regions share conserved centromere-specific satellite motifs with different copy numbers and structures. In addition, the similarity of *CentO* repeats in the same chromosome is higher than across chromosomes, supporting a model of local expansion and homogenization. Both genomes have over 395 non-TE genes located in centromere regions, of which ~41% are actively transcribed. Two large structural variants at the end of chromosome 11 affect the copy number of resistance genes between the two genomes. The availability of the two gap-free genomes lays a solid foundation for further understanding genome structure and function in plants and breeding climate-resilient varieties.

Key words: rice genome, ZS97, MH63, hybrid rice, centromere architecture

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., Zhou Y., Zuccolo A., Long E., Lee S., Talag J., Zhou R., Zhu X.-T., Yuan D., Udall J., Xie W., Wing R.A., Zhang Q., Poland J., Zhang J., and Chen L.-L. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant. **14**, 1757–1767.

INTRODUCTION

The *Oryza sativa* groups "*xian/indica*" and "*geng/japonica*" are two major types of Asian cultivated rice (Wang et al., 2018). The *xian* varieties contribute to over 70% of rice production worldwide and are genetically more diverse than *geng* rice. Over the past 30 years, two *xian* varieties, Zhenshan 97 (ZS97) and Minghui 63 (MH63), have emerged as important model systems in rice breeding and genomics, being the parents of the elite hybrid Shanyou 63, historically the most widely cultivated rice hybrid in China. Understanding the biological mechanisms behind the elite combination of ZS97 and MH63 to

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.

form the Shanyou 63 hybrid is foundational to help unravel the mystery of heterosis and for future advancements in breeding (Yu et al., 1997; Hua et al., 2002, 2003; Huang et al., 2006; Zhou et al., 2012). Further, ZS97 and MH63 represent two major varietal subgroups in *xian* rice, as they show many complementary agronomic traits, and a number of important genes have been cloned based on genetic populations generated using these two varieties as parents (Sun et al., 2004; Fan et al., 2006; Xue et al., 2008). Although we previously generated two reference genome assemblies, ZS97RS1 and MH63RS1, in 2016, approximately 10% of each genome remained unassembled or unplaced (Zhang et al., 2016a). Upon further analysis and editing we were able to fill the majority of gaps in each assembly and released upgraded versions of these two assemblies in 2018 (https://rice.hzau.edu.cn), yet eight (ZS97) and seven (MH63) gaps still remained. In recent years, several high-quality rice genomes, including Shuhui498 (Du et al., 2017) and two circum-basmati rice genomes (Choi et al., 2020), were reported, but no gapless genome has been obtained in rice or other plants up to now.

Centromeres are essential for maintaining the integrity of chromosomes during cell division and ensure the fidelity of their inheritance. Although centromeres are associated with a number of unique sequences, including a 155-bp centromere-specific satellite repeat, *CentO*, and a centromere-specific retrotransposon (Cheng et al., 2002), the functional centromere is defined by epigenetic replacement of histone H3 with the centromere-specific histone H3-like protein (CENH3) (Talbert et al., 2002). Thus, centromeres are not strictly a genetic feature of the genome. Unfortunately, until now, centromeres have remained largely underexplored, especially in larger genomes (Perumal et al., 2020), being notoriously difficult to completely assemble due to the highly repetitive sequence and complex structure. The limited examples of sequenced (gap-free) centromeres, such as chromosomes (chr) 4 and 8 of rice (Nagaki et al., 2004; Wu et al., 2004; Zhang et al., 2004) or chr2 and chr5 of maize (Wolfgruber et al., 2009), are known to be smaller with fewer repetitive sequences and, thus, possibly less representative of all centromeres (Kato et al., 2004; Nagaki et al., 2004). However, the centromeres that have been observed in their completeness have offered intriguing insights into centromere biology, including the presence of active genes (Nagaki et al., 2004), variable CENH3 density (Gent et al., 2015), and even epigenetic movement of the centromere to slightly or drastically different positions (Walkowiak et al., 2020). However, due to the challenges of studying complete centromeres, a global picture of the size, structure, and organization of centromeres has remained elusive. This is particularly challenging for comparison of centromere diversity in different genomes, which necessitates multiple gap-free assemblies.

In this study, we incorporated high-coverage and accurate long-read sequence data and multiple assembly strategies and bridged all remaining assembly gaps across each genome. These efforts resulted in two gap-free genome assemblies of *xian* rice varieties ZS97 and MH63, the first gap-free plant genome assemblies publicly available to date. The availability of the gap-free genome sequences provided the first opportunity for global analysis and comparison of the centromeres of all chromosomes side by side across both rice varieties. More than expected, at least 395 non-transposable element (non-TE) genes were identified in rice centromere regions, ~41% of which were found to be actively transcribed. The sequences and analyses have updated the view of the whole structure and function of the rice genome.

RESULTS AND DISCUSSION

Assembly and validation of gap-free reference genome sequences for ZS97 and MH63

To develop the assemblies, 56.73 Gb (~150× coverage) and 86.85 Gb (~230× coverage) of PacBio reads (including both HiFi and CLR modes) were generated for ZS97 and MH63, respectively, using the PacBio Sequel II platform (Supplemental Figure 1 and Supplemental Table 1). The PacBio HiFi and CLR reads were assembled separately with multiple *de novo* assemblers, including Canu (Koren et al., 2017), FALCON (Carvalho et al., 2016), and MECAT2 (Xiao et al., 2017) (see Methods), and then the assembled contigs were merged with the two upgraded assemblies using Genome Puzzle Master (GPM) (Zhang et al., 2016b) (Supplemental Tables 2 and 3). Finally, two gap-free reference genomes were produced, named ZS97RS3 and MH63RS3, which contained 12 chromosomes with total lengths of 391.56 Mb and 395.77 Mb, respectively (Figure 1A, Table 1). Compared with the previous bacterial artificial chromosome (BAC)-based RS1 genome assemblies, the new RS3 assemblies included ~36–45 Mb of additional sequence by filling 223 (ZS97RS1) and 167 (MH63RS1) gaps across both genomes (Supplemental Table 4). In addition, the new assemblies corrected some misoriented or misassembled regions caused by reliance on the Os-Nipponbare-Reference-IRGSP-1.0 sequence as a guide to produce the RS1 pseudomolecules (e.g., the 6-Mb inversion on chr6) (Supplemental Figure 2A–2C and Supplemental Table 4). These anomalies were corrected by newly assembled contigs that were long enough to span the previously ambiguous regions. Finally, using the seven-base telomeric repeat (CCCTAAA at the 5′ end or TTTAGGG at the 3′ end) as a sequence query, we identified 19 and 22 telomeres that resulted in 7 and 10 telomere-to-telomere pseudomolecules in the ZS97RS3 and MH63RS3 assemblies, respectively (Figure 1A and Supplemental Tables 5 and 6).

The accuracy and completeness of the RS3 assemblies were validated in multiple ways. Chromosome conformation capture sequencing (Hi-C) and Bionano optical maps showed high consistency across all pseudomolecules, demonstrating correct ordering and orientation (Supplemental Figure 3 and Supplemental Table 2). Genome completeness was demonstrated by high mapping rates with various raw sequences, such as raw PacBio HiFi and CLR, Illumina pair-end reads, paired BAC-end sequences, and paired-end short reads from 48 RNA-sequencing libraries, all of which mapped at over 99% across each assembly (Supplemental Tables 7–9). The evenly distributed breakpoints of aligned short and long reads indicated that the full genome was highly contiguous, with high accuracy at the single-base level in these final assemblies (Supplemental Figure 4). For gene content assessment, both ZS97RS3 and MH63RS3 assemblies captured 99.88% of a BUSCO 1614 reference gene set (Supplemental Table 10). Long terminal repeat (LTR) annotation further revealed the LTR assembly index for the ZS97RS3 and MH63RS3

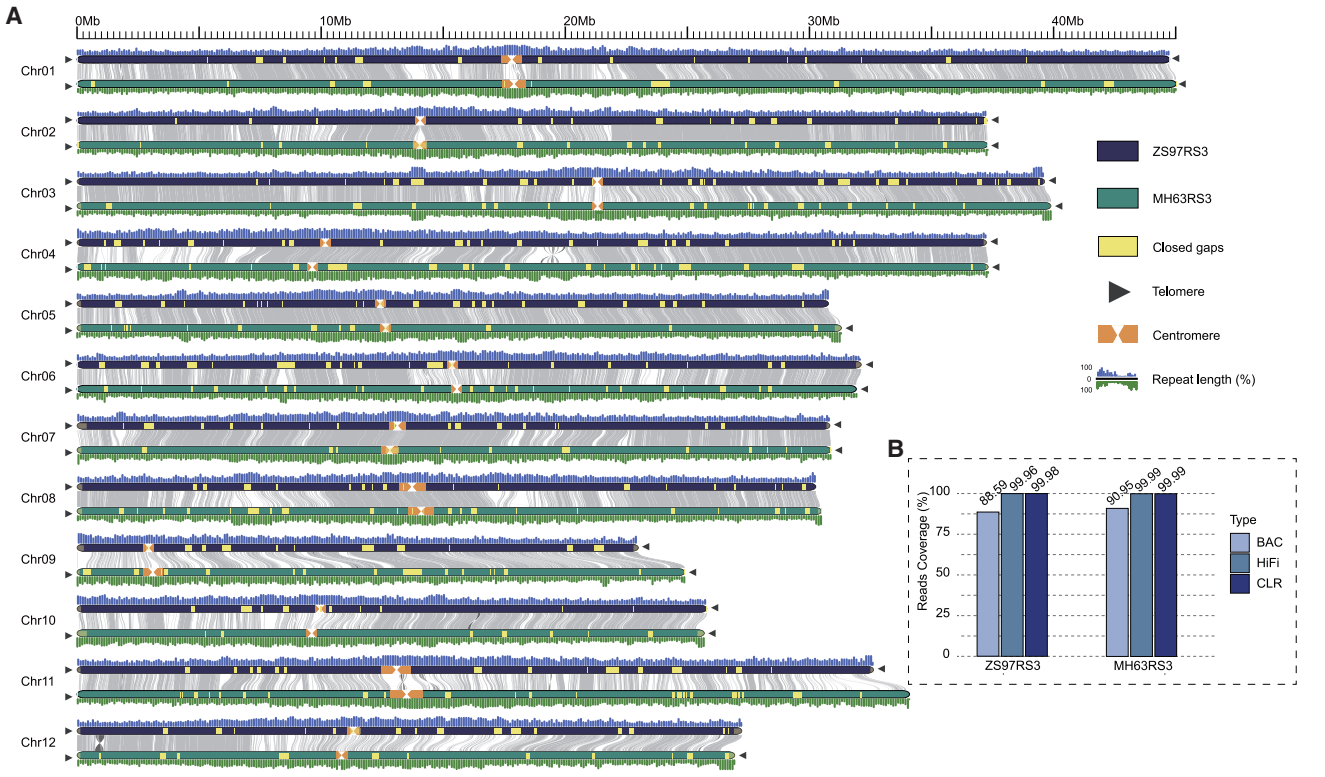

Figure 1. Two gap-free reference genomes of rice.

(A) Collinearity analysis between ZS97RS3 and MH63RS3. The collinear regions between ZS97RS3 and MH63RS3 are shown linked by gray lines. All the RS1 gap regions closed in RS3 are shown in yellow blocks. The black triangles indicate the presence of telomere sequence repeats. Repeat percentage distribution is plotted above or under each chromosome in 100-kb bins.

(B) Histogram showing the read coverage for different libraries in MH63RS3 and ZS97RS3, including BAC, HiFi, and CLR reads.

assemblies are 24.01 and 22.74, respectively, which meets the standard of gold/platinum reference genomes (Ou et al., 2018; Mussurova et al., 2020) (Table 1). More than 1500 rRNAs were identified in the ZS97RS3 and MH63RS3 assemblies (Supplemental Figure 5), whereas only tens were identified in the original RS1 assemblies.

The success of two rice gap-free reference genomes was achieved with a combination of deep-coverage sequence datasets from multiple platforms and cutting-edge technologies and assemblers. Sequence data generated by different sequencing technologies and library types are complementary to one another (Logsdon et al., 2020); in particular, PacBio HiFi sequencing with high accuracy on tens of kilobases-sized reads has provided great resources for the assembly of complex heterozygous regions and centromeres (Figure 1B). Likewise, every assembler has its own characteristics and strengths; thus the combination of multiple assembling tools can lead to better results. From our results, we would recommend using the assembly output from the latest HiCanu (Nurk et al., 2020) as backbone sequences and perform manual editing with integration of other *de novo* assemblers' outputs in the GPM pipeline (Zhang et al., 2016b). It is still intractable to fully assemble a gap-free genome without human intervention; as seen in our study, individual assemblers alone could not solve all the puzzles in a genome. At present, manual curation is necessary to handle chimeric contigs and collapsed repeats and correct errors in highly complicated regions (see Supplemental Note 1 for details).

Annotation and comparison of gap-free reference genome sequences for ZS97 and MH63

To annotate the ZS97 and MH63 RS3 assemblies for TEs and other repetitive sequences, we used RepeatMasker (Zhi et al., 2006) with the latest Repbase (Bao et al., 2015) and TIGR Oryza Repeat Database (v.3.3) (Ouyang and Buell, 2004) as libraries. As a result, we identified 465 242 TEs in ZS97RS3 (181.00 Mb in total length) and 468 675 TEs in MH63RS3 (~182.26 Mb) (Supplemental Tables 11 and 12), which accounted for ~46.16% and ~45.99% of each assembly and were approximately 5% greater than in the previous RS1 assemblies (i.e., ZS97RS1 = 41.28%; MH63RS3 = 41.58%). In addition to the updated repeat library, the repeat content increases were primarily a result of gaps being closed in TE-rich regions, with 82.9% of the 45 Mb closed-gaps in ZS97RS3 and 84.2% of the 36 Mb closed-gaps in MH63RS3 being in TEs.

Next, we employed MAKER-P (Campbell et al., 2014) to annotate the ZS97RS3 and MH63RS3 assemblies with the same evidence data used to annotate the RS1 assemblies (Supplemental Figure 1). To retain consistency across different assembly versions, 51 027 and 50 341 previously annotated gene models in the ZS97RS1 and MH63RS1 assemblies, respectively, were lifted onto the RS3 assemblies. Combining models annotated

Genomic feature	ZS97RS3	MH63RS3
Total size of assembled contigs (Mb)	391.562	395.765
Number of contigs (gaps)	12 (0)	12 (0)
Number of telomeres/subtelomeres	19	22
Number of centromeres	12	12
GC content (%)	43.61	43.64
Number of gene models/transcripts	60 935	59 903
Number of non-TE gene loci	39 258	39 406
Total size of TEs (Mb)	180.97	182.26
BUSCOs (%)	99.88	99.88
LTR assembly index score	24.01	22.74

Table 1. Characteristics of the ZS97RS3 and MH63RS3 genomes.

with MAKER-P in the newly assembled regions, the final annotations in ZS97RS3 and MH63RS3 contained 60 935 and 59 903 gene models, of which 39 258 and 39 406 were classified as non-TE gene loci (Table 1). This resulted in 4648 (ZS97) and 2082 (MH63) more non-TE genes, an increase of 11.8% and 5.3%, respectively, than previously identified in the RS1 assemblies. More than 92% of all annotated gene models were supported by homologies with known proteins or functional domains in *Oryza* and other species (Supplemental Tables 13 and 14).

Based on our new assemblies, the annotation and comparative analyses of non-coding RNAs (transfer RNAs, ribosomal RNAs, small nucleolar RNAs, microRNAs) (Supplemental Figure 5), single-nucleotide polymorphisms (SNPs), and insertions/deletions (indels) among ZS97, MH63, and Nipponbare (Supplemental Figure 6 and Supplemental Table 15); presence/absence variations (PAVs) (Supplemental Table 16); and genes in different categories (identical, same length, collinear, divergent, and variety-specific genes) (Supplemental Table 17) that were previously identified in the RS1 versions were updated.

After comparing the PAV distribution across each chromosome of both gap-free assemblies, we noticed an abundance of structural variations near the ends of the long arms of chromosome 11 (Figure 2A). Two large structural variations, one expansion region (30.75–31.57 Mb) and one insertion region (31.90–32.76 Mb), were uniquely detected in MH63 (hereafter named MH-E and MH-I, respectively). Raw sequencing read alignments to these two regions clearly showed that MH-E and MH-I could be continuously covered by MH63 reads but only partially covered by ZS97 reads (Supplemental Figure 7). Meanwhile, previous studies showed that nucleotide-binding site leucine-rich repeat (NLR) proteins were enriched in chromosome 11 (Rice Chromosomes 11 and 12 Sequencing Consortia, 2005). Hence, we performed a genome-wide homology search for NLR or NLR-like genes in both ZS97 and MH63 RS3 assemblies (Figure 2B). When combining the PAV and NLR(-like) distribution together, we could determine that both MH-E and MH-I regions have more NLR(-like) content than the corresponding region in the ZS97RS3 assembly (30.51–30.69 Mb and 30.88–30.94 Mb, respectively) (Supplemental Figure 7). In the MH-E region, most of the NLR(-like) genes in ZS97 amplified 2–10 times in MH63 (Figure 2C and Supplemental Table 18). Interestingly, these genes are more likely to be expressed in root than in other tissues (Figure 2C and Supplemental Figure 7C and Supplemental Table 18). In the 857-kb MH-I region, 11 NLR(-like) genes also had higher expression levels in roots than in other tissues (Figure 2D and Supplemental Table 19). We further scanned the MH-E and MH-I homologous regions in 25 additional high-quality reference genomes (Zhou et al., 2020), and unexpectedly, none of them had both complete MH-E and MH-I at the same time (Figure 2E and Supplemental Figure 8 and Supplemental Table 20). This unique genomic characteristic of MH63 could potentially explain, partially at least, its superior resistance to rice diseases (Chen, 2001).

Location and analyses of rice centromeres

To identify the location and sequence of functional centromeres in our gap-free genomes, we used the rice CENH3 antibody for chromatin immunoprecipitation and sequencing (ChIP-seq) of the captured DNA fragments (Figure 3A and 3B). To confirm the specificity of ChIP experiments, we used fluorescence *in situ* hybridization of ChIPed DNA on MH63 and ZS97 metaphase chromosomes, the results of which showed strong signals at the centromere for each chromosome, supporting the enrichment of centromeric sequences (Figure 3B).

Using the ChIP-seq mapping coverage in each genome, we used defined criteria to delimit the boundaries of each centromere and determined that the sizes of rice centromeres varied, e.g., from 0.6 Mb to 1.8 Mb in ZS97RS3 and from 0.8 Mb to 1.8 Mb in MH63RS3 (Supplemental Figure 9 and Supplemental Tables 21–22). We then classified rice centromeres into core and pericentromere regions. *CentO*-enriched regions (CoERs) were identified by sequence homology to the 155- to 165-bp *CentO* satellite repeats, all of which showed high levels of CENH3 binding (Cheng et al., 2002). Pericentromere regions were further delimited by enriched ChIP-seq signals. We manually checked the entire length of each centromere region, including 50-kb flanking regions of both boundaries, for both MH63RS3 and ZS97RS3 and found that the HiFi and CLR reads were evenly mapped with no observable within-read breakpoints (Figure 3C and Supplemental Figure 10), which provides strong evidence that each of the 12 centromeres in both gap-free reference genomes was contiguous and of high quality. The lengths of the CoERs varied almost 10-fold, ranging from 76 kb to 726 kb in different chromosomes in MH63RS3 (Supplemental Figure 9 and

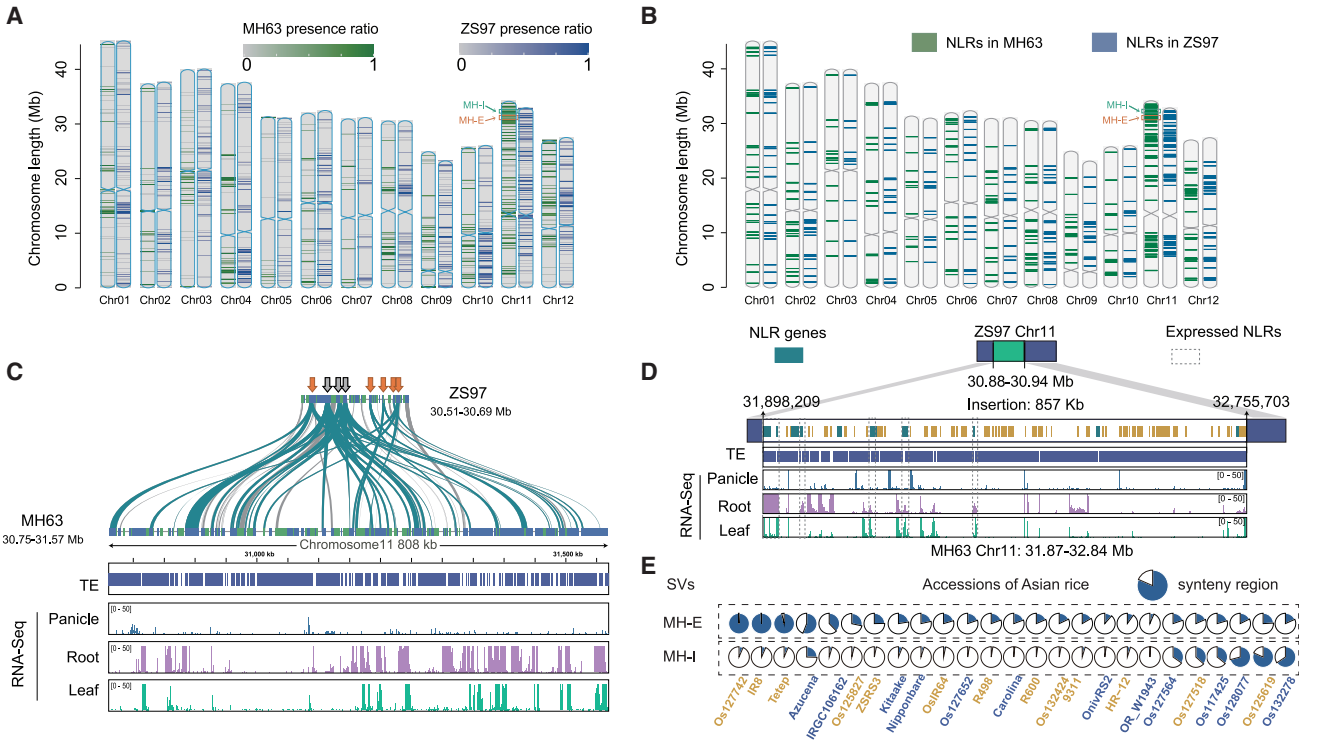

Figure 2. Structural variations of ZS97RS3 and MH63RS3 genomes.

(A) Distribution of the difference regions between ZS97RS3 and MH63RS3 chromosomes.

(B) Distribution of the NLR genes of ZS97RS3 and MH63RS3 on the chromosomes.

(C) The expansion structural variation MH-E in MH63RS3. The structure of MH-E at the end of chromosome 11 of MH63RS3. From top to bottom: the gene collinearity of ZS97RS3 and MH63RS3, the TE distribution, and the gene expression in this region.

(D) The insertion structural variation MH-I in MH63RS3. From top to bottom: the gene collinearity of ZS97RS3 and MH63RS3, the TE distribution, and the gene expression in this region.

(E) Coverage ratio of two structural variations (SVs; MH-E and MH-I) in 25 rice varieties.

Supplemental Table 22). There was also variable and non-uniform distribution of the CENH3 density, with some centromeres having CENH3 association almost exclusively on the *CentO* repeat (e.g., MH63 *Cen1* and *Cen2*), while other centromeres had CENH3 loading primarily outside of the *CentO* repeats (e.g. MH63 *Cen5*) (Figure 3A). Comparative analysis at the defined centromeres revealed relatively high synteny with conservation of genes and sequence order, but interesting differences in CENH3 density (e.g., chr4) (Supplemental Figure 11). We observed that centromere expansion corresponded to increased range of CENH3 (e.g., chr2) and that CENH3 loading traveled with original chromatin during duplications (e.g., chr10) and translocations/rearrangements (e.g., chr11). Notably, the centromere primary structure at the sequence level might not directly or fully reflect the architecture of a functional centromere, as we found that the ChIP-seq signals could be concentrated in a region with lots of structure variations (e.g., chr1) (Figure 3D), or the signals might vary, even in the centromeres that have very highly conversed sequence (e.g., chr4 and chr8) (Supplemental Figure 11). It is important to note that, while we have used consistent criteria to delimit the centromeres and observe clear differences in the size and structure of different centromeres, the exact boundaries of these epigenetic features of the genome maintain a level of subjectivity, just as the base and boundary of a mountain can vary depending on the defined criteria.

Analysis of all centromeres in both assemblies identified 395 and 539 non-TE genes in ZS97 and MH63, respectively. Of these non-TE genes, 163 in ZS97 (41.27%) and 235 in MH63 (43.60%) were found to be transcribed, which was much lower than the average gene transcription rate of the whole genome (>61%) (Supplemental Tables 23–26). In addition, 76.69% of the transcribed non-TE genes were expressed throughout the plant, including panicle, leaf, and root, and the proportion of specific expression was extremely low. This characterizes a constitutive expression of the non-TE genes in the centromere (Supplemental Table 26). Only two non-TE genes located in CoERs in ZS97 (18.18%) and MH63 (28.57%) were found to be transcribed, and most of the actively transcribed genes were located in the pericentromere regions (Supplemental Table 26). As an example of this gene distribution, MH63RS3 chr1 (~1.6 Mb) contained a 726-kb CoER composed of 3228 *CentO* sequences and only one non-TE gene, relative to the pericentromere regions, which contained 114 *CentO* sequences and 61 non-TE genes (Figure 3E and Supplemental Tables 22 and 23).

We detected 40 and 25 genes with low sequence identity in the centromere regions of ZS97 and MH63, respectively, which were not identified in other rice reference genomes. The results of PCR and RT-PCR showed that some of these genes are unique to MH63 and ZS97, such as OsMH_12G0168500 and OsZS_12G0181100, while others were missed due to incomplete

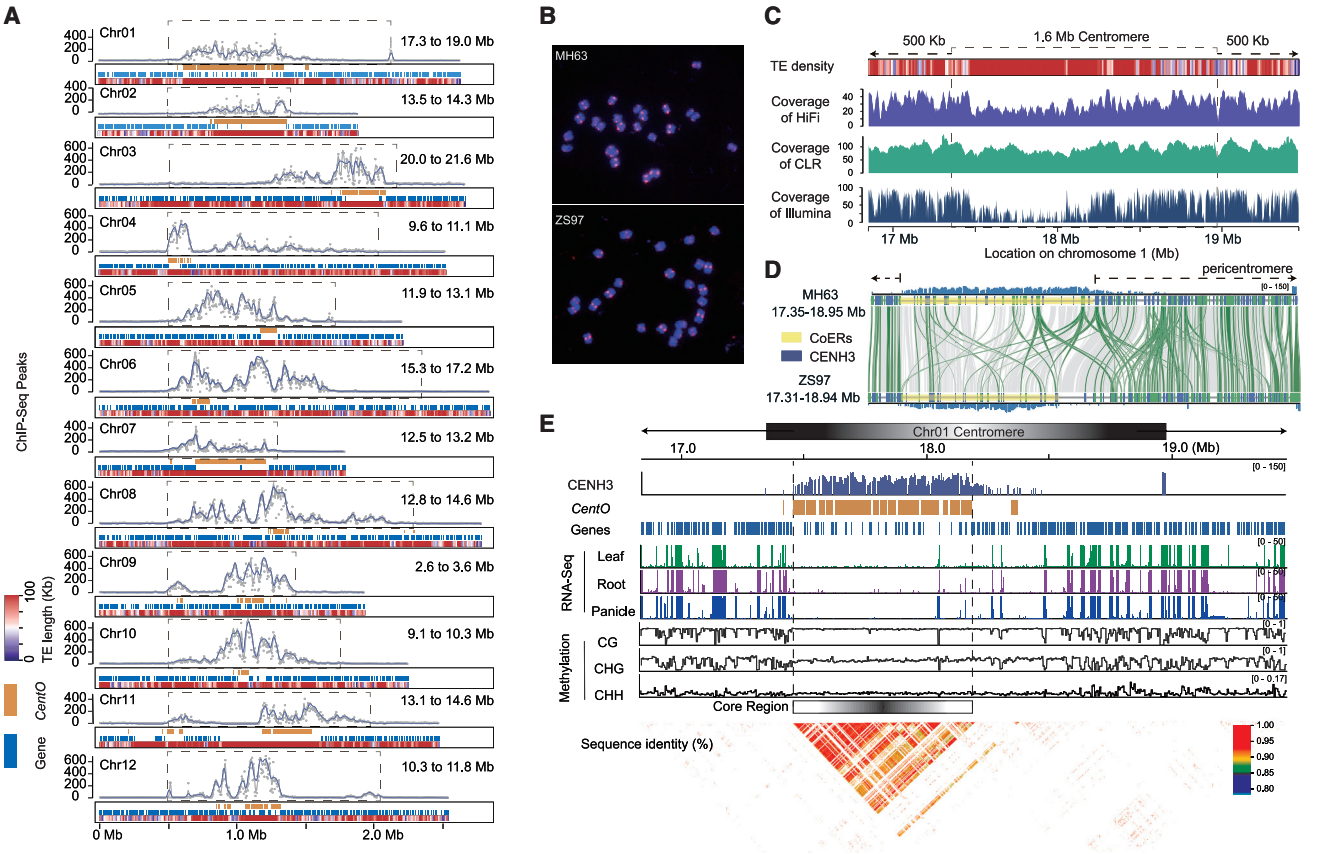

Figure 3. Characterization of complete rice centromeres.

(A) The delimiting of MH63RS3 centromeres. The layers of each chromosome graph indicate (1) the density of read mapping from CENH3 ChIP-seq with sliding windows of 10 kb and 20 kb shown in gray and blue lines, respectively; (2) the *CentO* satellite distribution; (3) non-TE gene distribution; and (4) TE distribution, respectively. The dotted frame represents the defined centromere region.

(B) Fluorescence *in situ* hybridization of mitotic metaphase chromosomes in MH63 and ZS97 using CENH3 ChIP-DNA as probe (red) with chromosomes counterstained with DAPI (blue).

(C) Coverage of HiFi, CLR, and Illumina reads and distribution of TEs in the centromere on chr1 (extended 500 kb left and right) of MH63RS3.

(D) The pairwise synteny visualization of chr1 centromere regions between ZS97RS3 and MH63RS3. Green lines link synteny genes between ZS97RS3 and MH63RS3. Yellow blocks are CoERs.

(E) Characteristics of the centromere on chr1 of MH63RS3. The 10 layers demonstrate the histone CENH3 distribution, *CentO* satellite distribution, gene distribution, gene expression level (in leaf, root, and panicle), methylation distribution (of CG, CHG, and CHH), and *CentO* sequence similarity, respectively.

assembly in other rice reference genomes (e.g., OsZS_12G0171400, OsZS_10G0147900, and OsMH_12G0165900). Interestingly, some homologous centromere genes were found to have structural variations in different rice varieties, which may affect the gene expression activity, such as OsZS_12G0181900 (Supplemental Figures 12 and 13). The gap-free genomes bring new opportunities for the identification of unique genes and large-effect structural variations in these "dark matter" regions.

Comparative analysis between the genomes revealed that 72% of the gene families were shared in the centromere regions of ZS97 and MH63 (Supplemental Figure 14), with up to ~91% conserved genes in the centromere of chromosome 1 (Figure 3D). These ratios were similar for other well-assembled chromosomes in other genomes (Supplemental Table 27). This conservation of genes could be extended throughout the population (K = 15) of cultivated Asian rice, in which the average rate of conserved genes is ~87%, especially across the chr5, chr9, and chr12 centromeres (Supplemental Table 27).

Gene ontology (GO) analysis showed that genes related to the GO terms "transcription from RNA polymerase III promoter," "nucleic acid binding," and "nucleoplasm part" were significantly enriched in ZS97 and MH63 centromere regions (Supplemental Figure 10B and 10C, Supplemental Tables 28 and 29). Overall, these GO terms tend to have similar functions (Supplemental Figure 15). However, GO terms among centromeres of different chromosomes of the same variety were very different, e.g., the average overlapping ratio was 37% in MH63 (Supplemental Tables 30 and 31). We also found that the methylation levels of CG and CHG in the centromeric regions were two-fold higher than that of the whole genome (Supplemental Table 32). This phenomenon was particularly prominent in *CentO* clustered regions.

In terms of the total sequence reads, we observed that the centromeric regions had slightly lower depth of mapped raw sequence reads than non-centromeric regions, which may be caused by highly repetitive elements; however, the lengths of those reads

in centromeric and non-centromeric regions were broadly consistent (Supplemental Figure 14). Detailed sequence analysis revealed abundant TEs in the centromeric regions accounting for 78%–80% of the functional centromeres (Supplemental Tables 33 and 34). In particular, the proportion of LTR/gypsy TEs account for over 90% of the repetitive sequences (Supplemental Figure 14), which is an obvious barrier to fully assembling a centromere region.

To better understand the long-range organization and evolution of the CoERs, we generated a heatmap showing pairwise sequence identity of 1 kb along the centromeres (Supplemental Figure 16A), and observed that the *CentO* sequences had the highest similarity in the middle and declined to both sides (Supplemental Figure 16A). Furthermore, the profile of *CentO* sequences (Supplemental Figure 16B) illustrated the conservation of rice centromeres on the genomic level. We constructed a phylogenetic tree by using all the 155- to 165-bp *CentO* repeats from MH63RS3 and ZS97RS3, and observed that *CentO* satellite repeats from chr1, chr2, chr4, chr5, chr7, chr11, and chr12 of the two genomes were clustered into seven distinct branches (Supplemental Figure 17), indicating that the similarity of *CentO* between the two genomes on homologous chromosomes is higher than the similarity across chromosomes, supporting models of repeated amplification events involving the central domain and local homogenization (Lee et al., 2006).

To determine if the centromere architecture found in ZS97 and MH63 was conserved among other Asian rice accessions, we compared the ZS97/MH63 CoER sequences with 15 high-quality PacBio genome assemblies that represent the population structure of cultivated Asian rice (Zhou et al., 2020). The results revealed that lengths of *CentO* satellite repeats in the CoERs of the same chromosomes varied significantly between varieties within the same subspecies (or natural populations) of Asian rice (Supplemental Tables 35 and 36).

The gap-free assemblies produced here enabled the first global assessment of functional centromeres in plants. The large 10-fold variation in the number and distribution of centromeric repeats across the different chromosomes and between the genomes gives a detailed picture of the large amount of centromeric diversity both within and among plant genomes (Cheng et al., 2002). Centromeric regions, while critical for fidelity and segregation of chromosomes, are largely inaccessible to breeding due to greatly reduced recombination (Chen et al., 2002), particularly in larger genomes (The International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018). The detailed understanding of centromere architecture and gene content, therefore, affords insight into the challenge of developing favorable allele combinations in the absence of natural recombination, using hybrid complementation or gene editing, or even precisely inducing recombination. The large number of genes, as well as the relatively high conservation of genes and gene families, in the functional centromere gives a better foundation for understanding the mechanisms of heterosis, which is often associated with complementation of divergent pericentromeric regions of hybrid parents (Zhou et al., 2012; Thiemann et al., 2014).

In conclusion, the generation and validation of two gap-free assemblies of ZS97 and MH63, presented here, provide a clear picture of the primary sequence architecture of the *xian/indica* rice genomes. Such resources will serve to develop a fundamental and comprehensive model for the study of heterosis, and other basic and applied research, and pave the way toward a new standard for assembling reference genomes of other plant species.

METHODS

Plant materials and sequencing

Fresh young leaf tissue was collected from *O. sativa* ZS97 and MH63 plants. We constructed SMRTbell libraries as described in a previous study (Pendleton et al., 2015). The genomes of MH63 and ZS97 were sequenced using the PacBio Sequel II platform (Pacific Biosciences), to produce 8.34 Gb HiFi reads (~23× coverage) and 48.39 Gb CLR reads (~131× coverage), for the ZS97 genome, and 37.88 Gb HiFi reads (~103× coverage) and 48.97 Gb CLR reads (~132× coverage) for the MH63 genome.

The Truseq Nano DNA HT sample preparation kit was used, following the manufacturer's standard protocol (Illumina), to generate the libraries for Illumina paired-end genome sequencing. These libraries were sequenced using the Illumina HiSeq X Ten platform to generate 150-bp paired-end reads with 350-bp insert size and produce 25 Gb reads (~69× coverage) for ZS97 and 28 Gb reads (~76× coverage) for MH63.

Plant tissues used for optical mapping were extracted using the Bionano plant tissue extraction protocol (Staňková et al., 2016). Extracted DNA was embedded in Bio-Rad LE agarose for subsequent washes with Tris-EDTA and for proteinase K (0.8 mg/ml) and RNase A (20 μl/ml) treatments in lysis buffer. The agarose plugs were then melted using agarase (0.1 U/μl, New England Biolabs) and dialyzed on Millipore membranes (0.1 μm) with Tris-EDTA to equilibrate ion concentrations. The DNA was then nicked with the nickase restriction enzyme BssSI (2 U/μl). Labeled nucleotides were incorporated at breakpoints and the DNA was counterstained. Each sample was loaded onto two nanochannel flow cells of a Bionano Irys machine for DNA imaging.

Genome assembly and assessment

Seven tools based on different algorithms were used to assemble the genomes of ZS97 and MH63: (1) Canu v.1.8 (Koren et al., 2017) was used to assemble the genomes with default parameters; (2) FALCON toolkit v.0.30 (Carvalho et al., 2016) was applied for assembly with the parameters pa_DBsplit_option = -s200 -x500, ovlp_DBsplit_option = -s200 -x500, pa_REPmask_code = 0,300;0,300;0,300, genome_size = 400000000, seed_coverage = 30, length_cutoff = -1, pa_HPCdaligner_option = -v -B128 -M24, pa_daligner_option = -k18 -w8 -h480 -e.80 -l5000 -s100, falcon_sense_option = –output-multi –min-idt 0.70 –min- cov 3 –max-n-read 400, falcon_sense_greedy = False, ovlp_HPCdaligner_option = -v -M 24 -l500, ovlp_daligner_option = -h60 -e0.96 -s1000, overlap_filtering_setting = –max-diff 100 –max-cov 100- -min-cov 2, length_cutoff_pr = 1000; (3) MECAT2 (Xiao et al., 2017) was utilized to assemble with the parameters GENOME_SIZE = 400000000, MIN_READ_LENGTH = 2000, CNS_OVLP_OPTIONS = "", CNS_OPTIONS = "-r 0.6 -a 1000 -c 4 -l 2000", CNS_OUTPUT_COVERAGE = 30, TRIM_OVLP_OPTIONS = "-B", ASM_OVLP_OPTIONS = "-n 100 -z 10 -b 2000 -e 0.5 -j 1 -u 0 -a 400", FSA_OL_FILTER_OPTIONS = "–max_overhang = -1 –min_identity = - 1", FSA_ASSEMBLE_OPTIONS = "", GRID_NODE = 0, CLEANUP = 0, USE_GRID = false; (4) Flye release 2.6 (Kolmogorov et al., 2019) was set with "–genome-size 400m"; (5) Wtdbg2 2.5 (Ruan and Li., 2020) was used to assemble with parameters "-x sq, -g 400m" and then Minimap2 (Li, 2018) was employed to map the PacBio CLR data to the assembly results, and wtpoa was utilized to polish and correct the wtdbg2 assembly results; (6) NextDenovo v.2.1-beta.0 (https://github.com/Nextomics/NextDenovo) was applied for assembly with parameters "task = all, rewrite = yes, deltmp = yes, rerun = 3, input_type = raw, read_cutoff = 1k, seed_cutoff = 44382, blocksize =

2g, pa_correction = 20, seed_cutfiles = 20, sort_options = -m 20g -t 10 -k 40, minimap2_options_raw = -x ava-ont -t 8, correction_options = -p 10, random_round = 20, minimap2_options_cns = -x ava-pb -t 8 -k17- w17, nextgraph_options = -a 1"; (7) Miniasm-0.3-r179 (Li, 2016) was used with default parameters.

Based on the results of these seven software tools, GPM (Zhang et al., 2016b) was then used to integrate and optimize the assembled contigs and visualize complete chromosomes. Based on the HiFi and CLR sequencing data, we used the GenomicConsensus package of SMRTLink/7.0.1.66975 (https://www.pacb.com/support/) to polish the assembled genomes twice with the Arrow algorithm, using the parameter –algorithm=arrow. Pilon (Walker et al., 2014) was used for polishing the genomes based on Illumina data with the parameters –fix snps, indels. This process was repeated twice. Molecules were then assembled using the Bionano IrysSolve pipeline (https://bionanogenomics.com/support-page/) to create optical maps. Images were interpreted quantitatively using Bionano AutoDetect 2.1.4.9159 and data were visualized using IrysView v.2.5.1. These assemblies were used with draft genome assemblies to validate and scaffold the sequences. Bionano optical map data were aligned to the merged contigs using RefAlignerAssembler in the IrysView software package to perform the verification.

ZS97RS3 and MH63RS3 genome completeness was assessed using BUSCO v.4.0.6, which contained 1614 genes in the "embryophyta_odb10" dataset (Simão et al., 2015), with default parameters. In addition, we mapped the PacBio HiFi reads and PacBio CLR reads with Minimap2 (Li, 2018), Illumina reads with BWA-0.7.17 (Jo and Koh, 2015), BES/BAC reads with BLASTN v.2.7.1 (Altschul et al., 1990), Hi-C reads with HiC-Pro v.2.11.1 (Servant et al., 2015), and RNA-sequencing reads with Hisat2 v.2.1.0 (Kim et al., 2015) to both genome assemblies.

Gene and repeat annotations

MAKER-P (Campbell et al., 2014) version 3 was used to annotate the ZS97RS3 and MH63RS3 genomes. All evidence was the same as that used for RS1 genome annotations. To ensure consistency with the RS1 versions, genes that mapped in their entirety to the RS3 genomes were retained. New genes in gap regions were obtained from MAKER-P results (Campbell et al., 2014). Genes encoding transposable elements were identified and transitively annotated by searching against the MIPS-REdat Poaceae version 9.3p (Nussbaumer et al., 2013) database using TBLASTN (Altschul et al., 1990) with an E value of 1e−10. tRNAs were identified with tRNAscan-SE (Lowe and Eddy, 1997) using default parameters; rRNA genes were identified by searching each assembly against the rRNA sequences of Nipponbare using BLASTN v.2.7.1 (Altschul et al., 1990); microRNAs and small nuclear RNAs were predicted using INFERNAL of Rfam (Griffiths-Jones et al., 2005) (v.14.1). Repeats in the genome were annotated using RepeatMasker (Zhi et al., 2006) with RepBase (Bao et al., 2015), and TIGR Oryza Repeats (v.3.3) with the RMBlast search engine. For overlapping repeats in different classes, LTR retrotransposons were kept first, next terminal inverted repeats (TIRs), and then short and long interspersed nuclear elements, and finally helitrons. This priority order was based on stronger structural signatures. In addition, the known nested insertion models (LTR into helitron, helitron into LTR, TIR into LTR, LTR into TIR) were retained. The identified repetitive elements were further characterized and classified using PGSB repeat classification schema. LTR_FINDER (Xu and Wang 2007) was used to identify complete LTR-RTs with target site duplications, primer binding sites, and polypurine tracts.

Chromatin immunoprecipitation and ChIP-seq

Procedures for ChIP were adopted from Nagaki et al. (2003) and Walkowiak et al. (2020) using nuclei from 4-week-old seedlings. Chromatin with the nuclei was digested with micrococcal nuclease (Sigma-Aldrich, St. Louis, MO) to liberate nucleosomes. For ChIP, we used anti-centromeric histone 3 antibody (N terminus), which reacts with 18.5 kDa CenH3 protein from *O. sativa* purchased from Antibodies-online (Limerick, PA; cat. no. ABIN1106669). The digested mixture was then incubated overnight with 3 μg of rice CENH3 antibody at 4°C. The target antibodies were then captured from the mixture using Dynabeads Protein G (Invitrogen, Carlsbad, CA). ChIP-seq libraries were then constructed using a TruSeq ChIP library preparation kit (Illumina, San Diego, CA) following the manufacturer's instructions, and the libraries were sequenced (2 × 150 bp) on an Illumina HiSeq X Ten.

Fluorescence *in situ* hybridization

Slide preparation

Mitotic chromosomes were prepared as described by Koo and Jiang (2009) with minor modifications. Root tips were collected from plants and treated in a nitrous oxide gas chamber for 1.5 h. The root tips were fixed overnight in ethanol:glacial acetic acid (3:1) and then squashed in a drop of 45% acetic acid.

Probe labeling and detection

The ChIPed DNAs were labeled with digoxigenin-16-dUTP using a nick-translation reaction. The clone, maize 45S rDNA (Koo and Jiang 2009), was labeled with biotin-11-dUTP (Roche, Indianapolis, IN). Biotin- and digoxigenin-labeled probes were detected with Alexa Fluor 488 streptavidin antibody (Invitrogen) and rhodamine-conjugated anti-digoxigenin antibody (Roche), respectively. Chromosomes were counterstained with DAPI in Vectashield antifade solution (Vector Laboratories, Burlingame, CA). Images were captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy, Thornwood, NY) using a cooled CCD camera (CoolSNAP HQ2, Photometrics, Tucson, AZ) and AxioVision 4.8 software. The final contrast of the images was processed using Adobe Photoshop CS5 software.

The completeness of centromeres on MH63RS3 and ZS97RS3 chromosomes

Based on the final RS3 genome assemblies, we used BLAST (Altschul et al., 1990) to align the *CentO* satellite repeats in rice to each reference genome with an E-value of 1e−5, and then use BEDtools (Quinlan, 2014) to merge the results with the parameter "-d 50000". If a region contained more than 10 consecutive *CentO* repeats with lengths longer than 10 kb, it was classified as a CoER.

For the identification of the whole centromere region, we used BWA-0.7.17 (Jo and Koh., 2015) to align the CENH3 ChIP-seq reads to MH63RS3 and ZS97RS3 genomes, and used SAMtools (Li et al., 2009) to filter the results (mapQ value <30) and count aligned reads in 10-kb bins; then we used MACS2 (Zhang et al., 2008) to call the peaks of CENH3 with parameters "-g dm -n chip –broad –broad-cutoff 0.1". Using the ChIP-seq mapping coverage as a signal indicator, we first manually delineated a range for each centromere as the approximate region to have CENH3 enrichment and the *CentO* satellite repeats. The boundary was then defined as the position where three or more consecutive ChIP signals remained >30. Then, we extended to both flanking sides (500 kb as the window size) until no ChIP-seq signal was found. The final positions, therefore, were delimited as start and end points of each chromosome centromere region. See details in Supplemental Figure 9 and Supplemental Tables 21 and 22 for our manual delimiting information and exact criteria of all ZS97 and MH63 centromeres.

To compare *CentO* sequence similarity, we first used BEDtools (Quinlan, 2014) to obtain sequences of centromere core regions, and divided them into 1-kb continuous sequences; then we used Minimap2 (Li, 2018) to align the sequences with the parameters "-f 0.00001 -t 8 -X –eqx -ax ava -pb"; and, finally, we used a custom Python script to filter the results file, and used R to generate a heatmap showing pairwise sequence identity (Logsdon et al., 2020).

PCR for gene verification

The DNA and RNA of MH63, ZS97, 9311, and Nipponbare were taken 15 days after germination. Genomic DNA was extracted using the CTAB

method, RNA was extracted with TransZol (TransGen, cat. no. ET101-01), and Hifair III 1st Strand cDNA Synthesis SuperMix for qPCR (Yeasen, cat. no. 11141ES10) was used for reverse transcription. PCR was performed using KOD FX (TOYOBO, F0935K).

Construction of phylogenetic tree of *CentO* monomers

Using a custom Python script, we extracted all 155-bp *CentO* sequences of MH63RS3 and ZS97RS3. The *CentO* monomers were then aligned by ClustalW (Chenna et al., 2003), and phylogenetic analysis was performed by the neighbor-joining method with a bootstrap value of 1000, and visualized with iTOL (Ivica et al., 2019). OrthoMCL (Li et al., 2003) was used to cluster the *CentO* sequences from chromosomes 1, 2, 7, and 11 in MH63RS3.

Telomere sequence identification

The telomere sequence 5′-CCCTAAA-3′ and the reverse complement of the seven bases were searched directly. In addition, we used BLAT (Kent, 2002) to search telomere-associated tandem repeat sequences from the TIGR *Oryza* Repeat database (Ouyang and Buell, 2004) in the whole genome.

Identification of PAVs between ZS97RS3 and MH63RS3

The ZS97RS3 assembly was aligned to the MH63RS3 assembly using Mummer (4.0.0beta2) (Marçais et al., 2018) with parameters settings "-c 90 -l 40". Then we used the "delta-filter -1" parameter with the one-to-one alignment block option to filter the alignment results. Further, "show-diff" was used to select for unaligned regions as the PAVs.

Prediction of NLR genes

We first predicted domains of genes with InterProScan (Jones et al., 2014), which can analyze peptide sequences against InterPro member databases, including ProDom, PROSITE, PRINTS, Pfam, PANTHER, SMART, and Coils. Pfam and Coils were used to prediction NLRs, which were required to contain at least one NB, TIR, or CC_R (RPW8) using the following reference sequences: NB (Pfam accession PF00931), TIR (PF01582), RPW8 (PF05659), LRR (PF00560, PF07725, PF13306, PF13855) domains, or CC motifs (Van de Weyer et al., 2019).

Identification of collinear orthologs

MCscan (Python version) (Tang et al., 2008) was used to identify collinear orthologs between chromosomes 11 of the ZS97RS3 and MH63RS3 genomes with default parameters.

ACCESSION NUMBERS

All the raw sequencing data generated for this project are archived at NCBI under accession nos. SRR13280200, SRR13280199, and SRR13288213 for ZS97, and SRX6957825, SRX6908794, SRX6716809, and SRR13285939 for MH63. The genome assemblies are available at NCBI (CP056052–CP056064 for ZS97RS3, CP054676–CP054688 for MH63RS3) or the National Genomics Data Center under BioProject no. PRJCA005549, and annotations are visualized with Gbrowse at http://rice.hzau.edu.cn. All the materials in this study are available upon request.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

This research was supported by the National Key Research and Development Program of China (2016YFD0100904 and 2016YFD0100802), the National Natural Science Foundation of China (31871269), the Hubei Provincial Natural Science Foundation of China (2019CFA014), and Fundamental Research Funds for the Central Universities (2662020SKPY010 to J.Z.).

AUTHOR CONTRIBUTIONS

L.-L.C., J.Z., R.W., and Q.Z. designed studies and contributed to the original concept of the project. J.P. and D.-H.K. performed the ChIP-seq and fluorescence *in situ* hybridization experiments. D.K., E.L., S.L., J.T., D.Y., J.U., and R.W. performed the genome and Bionano sequencing. J.-M.S., W.-Z.X., S.W., Y.-X.G., C.G., Y.H., J.-W.F., W.Z., Y.Z., A.Z., R.Z., and X.T.Z. performed genome assembling and annotation, comparative genomics analysis, and other data analysis. J.-M.S., W.-Z.X., S.W., J.P., D.-H.K., L.-L.C., and J.Z. wrote the paper. W.X., R.W., and Q.Z. contributed to revisions.

ACKNOWLEDGMENTS

We sincerely thank (1) Pacific Biosciences of California, Inc., for sequencing of MH63; (2) Wuhan Frasergen Bioinformatics Co., Ltd., for sequencing of ZS97; (3) the computing platform of the National Key Laboratory of Crop Genetic Improvement in HZAU for providing the computational resources; and (4) Dr. Jiming Jiang at MSU for his critical comments and constructive suggestions on our centromere analyses. No conflict of interest declared.

Received: May 8, 2021
Revised: June 16, 2021
Accepted: June 22, 2021
Published: June 22, 2021

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.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**:11.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. **164**:513–524.

Carvalho, A.B., Dupim, E.G., and Goldstein, G. (2016). Improved assembly of noisy long reads by k-mer validation. Genome Res. **26**:1710–1720.

Chen, H. (2001). Population Structure of Pyricularia Grisea from Central and Southern China and Comparative Mapping of QTL for Blast- and Bacterial Blight-Resistance in Rice and Barley (in Chinese), PhD Dissertation (Wuhan, China: Huazhong Agriculture University), pp. 77–78.

Chen, M., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., et al. (2002). An integrated physical and genetic map of the rice genome. Plant Cell **14**:537–545.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. **31**:3497–3500.

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.

Choi, J.Y., Lye, Z.N., Groen, S.C., Dai, X., Rughani, P., Zaaijer, S., Harrington, E.D., Juul, S., and Purugganan, M.D. (2020). Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. Genome Biol. **21**:21.

Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Li, Y., Zhao, X., et al. (2017). Sequencing and de novo assembly of a near complete indica rice genome. Nat. Commun. **8**:15324.

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., and Zhang, Q. (2006). GS3, a major QTL for grain length and weight and minor QTL for

grain width and thickness in rice, encodes a putative transmembrane protein. Theor. Appl. Genet. **112**:1164–1171.

Gent, J.I., Wang, K., Jiang, J., and Dawe, R.K. (2015). Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres. Genetics **200**:1105–1116.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. **33**:D121–D124.

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S., and Zhang, Q. (2003). Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U S A **100**:2574–2579.

Hua, J.P., Xing, Y.Z., Xu, C.G., Sun, X.L., Yu, S.B., and Zhang, Q. (2002). Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics **162**:885–1895.

Huang, Y., Zhang, L., Zhang, J., Yuan, D., Xu, C., Li, X., Zhou, D., Wang, S., and Zhang, Q. (2006). Heterosis and polymorphisms of gene expression in an elite rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. Plant Mol. Biol. **62**:579–591.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research. **47**:W256–W259.

Jo, H., and Koh, G. (2015). Faster single-end alignment generation utilizing multi-thread for BWA. Biomed. Mater. Eng. Suppl. **1**:S1791–S1796.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

Kato, A., Lamb, J.C., and Birchler, J.A. (2004). Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc. Natl. Acad. Sci. U S A **101**:13554–13559.

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. **12**:656–664.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. **37**:540–546.

Koo, D.H., and Jiang, J.M. (2009). Super-stretched pachytene chromosomes for fluorescence in situ hybridization mapping and immunodetection of cytosine methylation. Plant J. **59**:509–516.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**:722–736.

Lee, H.R., Neumann, P., Macas, J., and Jiang, J. (2006). Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. Mol. Biol. Evol. **23**:2505–2520.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics **32**:2103–2110.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**:3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Supgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079.

Logsdon, G.A., Vollger, M.R., Hsieh, P.H., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2020). The structure, function, and evolution of a complete human chromosome 8. Nature **593**:101–107.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. Plos Comput. Biol. **14**:e1005944.

Mussurova, S., Al-Bader, N., Zuccolo, A., and Wing, R.A. (2020). Potential of platinum standard reference genomes to exploit natural variation in the wild relatives of rice. Front Plant Sci. **11**:579980.

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.

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.

Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. **30**:1291–1305.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., and Spannagl, M. (2013). MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. **41**:D1144–D1151.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**:e126.

Ouyang, S., and Buell, C.R. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. **32**:D360–D363.

Pendleton, M., Sebra, R., Pang, A.W., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods **12**:780–786.

Perumal, S., Koh, C.S., Jin, L., Buchwaldt, M., Higgins, E.E., Zheng, C., Sankoff, D., Robinson, S.J., Kagale, S., Navabi, Z.K., et al. (2020). A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral Brassica genome. Nat. Plants **6**:929–941.

Quinlan, A.R. (2014). BEDTools: the swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics **47**:11.12.134.

Rice Chromosomes 11 and 12 Sequencing Consortia. (2005). The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. BMC Biol. **3**:20.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods **17**:155–158.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. **16**:259.

Simão, 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.

Staňková, H., Hastie, A.R., Chan, S., Vrána, J., Tulpová, Z., Kubaláková, M., Visendi, P., Hayashi, S., Luo, M., Batley, J., et al. (2016). BioNano genome mapping of individual chromosomes

supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol. J. **14**:1523–1531.

Sun, X., Cao, Y., Yang, Z., Xu, C., Li, X., Wang, S., and Zhang, Q. (2004). *Xa26*, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. Plant J. **37**:517–527.

Talbert, P.B., Masuelli, R., Tyagi, A.P., Comai, L., and Henikoff, S. (2002). Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell **14**:1053–1066.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science **320**:486–488.

The International Wheat Genome Sequencing Consortium (IWGSC), Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C.J., Choulet, F., Distelfeld, A., Poland, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**:eaar7191.

Thiemann, A., Fu, J., Seifert, F., Grant-Downton, R.T., Schrag, T.A., Pospisil, H., Frisch, M., Melchinger, A.E., and Scholten, S. (2014). Genome-wide meta-analysis of maize heterosis reveals the potential role of additive gene expression at pericentromeric loci. BMC Plant Biol. **14**:88.

Van de Weyer, A.-L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K., Jones, J.D.G., Dangl, J.L., Weigel, D., and Bemm, F. (2019). A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. Cell **178**:1260–1272.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature **588**:277–283.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557**:43–49.

Wolfgruber, T.K., Sharma, A., Schneider, K.L., Albert, P.S., Koo, D.H., Shi, J., Gao, Z., Han, F., Lee, H., Xu, R., et al. (2009). Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. Plos Genet. **5**:e1000743.

Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., Shibata, M., Ito, Y., Nakamura, M., Sakaguchi, M., Yoshihara, R., et al. (2004). Composition and structure of the centromeric region of rice chromosome 8. Plant Cell **16**:967–976.

Xiao, C.L., Chen, Y., Xie, S.Q., Chen, K.N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods **14**:1072–1074.

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., Zhou, H., Yu, S., Xu, C., Li, X., et al. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat. Genet. **40**:761–767.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q., and Saghai Maroof, M.A. (1997). Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. USA **94**:9226–9231.

Zhang, J., Chen, L.L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li, W., Song, J.M., Xie, W., et al. (2016a). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proc. Natl. Acad. Sci. USA **113**:E5163–E5171.

Zhang, J., Kudrna, D., Mu, T., Li, W., Copetti, D., Yu, Y., Goicoechea, J.L., Lei, Y., and Wing, R.A. (2016b). Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. Bioinformatics **32**:3058–3064.

Zhang, Y., Huang, Y., Zhang, L., Li, Y., Lu, T., Lu, Y., Feng, Q., Zhao, Q., Cheng, Z., Xue, Y., et al. (2004). Structural features of the rice chromosome 4 centromere. Nucleic Acids Res. **32**:2023–2030.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. **9**:R137.

Zhi, D., Raphael, B.J., Price, A.L., Tang, H., and Pevzner, P.A. (2006). Identifying repeat domains in large genomes. Genome Biol. **7**:R7.

Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J., and Zhang, Q. (2012). Genetic composition of yield heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. USA **109**:15847–15852.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., et al. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci. Data **7**:113.