The gap-free rice genomes provide insights for centromere structure and function exploration and graph-based pan-genome construction

The Gap-free Rice Genomes Provide Insights for Centromere Structure and Function Exploration and Graph-based Pan-genome Construction

Jia-Ming Song[1,2,#], Wen-Zhao Xie[1,#], Shuo Wang[1,#], Yi-Xiong Guo[1], Jesse Poland[3], Dal-Hoe Koo[3], Dave Kudrna[4], Evan Long[5], Yicheng Huang[1], Jia-Wu Feng[1], Wenhui Zhang[1], Seunghee Lee[4], Jayson Talag[4], Run Zhou[1], Xi-Tong Zhu[1], Daojun Yuan[1], Joshua Udall[5], Weibo Xie[1], Rod A. Wing[4,6,7], Qifa Zhang[1], Jianwei Zhang[1,*], 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]Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

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

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

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

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

[#] These authors contributed equally to this work.

[*]**Correspondence: Jianwei Zhang** (jzhang@mail.hzau.edu.cn),

Ling-Ling Chen (llchen@mail.hzau.edu.cn)

ABSTRACT

Asia rice (*Oryza sativa*) is divided into two subgroups, *indica/xian* and *japonica/geng*, the former has greater intraspecific diversity than the latter. Here, for the first time, we report the assemblies and analyses of two gap-free *xian* rice varieties 'Zhenshan 97 (ZS97)' and 'Minghui 63 (MH63)'. Genomic sequences of these elite hybrid parents express extensive difference as the foundation for studying heterosis. Furthermore, the gap-free rice genomes provide global insights to investigate the structure and function of centromeres in different chromosomes. All the rice centromeric regions share conserved centromere-specific satellite motifs but with different copy numbers and structures. Importantly, we show that there are >1,500 genes in centromere regions and ~16% of them are actively expressed. Based on MH63 gap-free reference genome, a graph-based rice pan-genome (Os-GPG) was constructed containing presence/absence variations of 79 rice varieties. Compared with the other rice varieties, MH63 contained the largest number of resistance genes. The acquisition of ZS97 and MH63 gap-free genomes and graph-based pan-genome of rice lays a solid foundation for the study of genome structure and function in plants.

Key words: gap-free genomes, ZS97, MH63, centromere structure, graph-based pan-genome

INTRODUCTION

Oryza sativa '*indica/xian*' and '*japonica/geng*' groups (in place of subsp. *indica* and subsp. *japonica* respectively) are two major groups of Asian cultivated rice (Wang et al., 2018). *Xian* rice varieties are broadly studied as they contribute over 70% of rice production worldwide and genetically more diverse than *japonica* rice. Over the past 30 years, two *xian* varieties Zhenshan 97 (ZS97) and Minghui 63 (MH63), combined with their elite hybrid Shanyou 63 (SY63), have been used as a research model in a series of fundamental studies due to three important facts: 1) ZS97 and MH63 represent two major varietal subgroups in *xian* rice, contain a number of important agronomic traits; 2) SY63 has historically been the most widely cultivated hybrid rice in China; 3) Understanding the biological mechanisms behind the elite combination of ZS97 and MH63 to form the SY63 hybrid is foundational to help unravel the mystery of heterosis which puzzled scientists for more than a century (Hua et al., 2002; Hua et al., 2003; Huang et al., 2006; Zhou et al., 2012). Although we previously generated two reference genome assemblies ZS97RS1 and MH63RS1 in 2016, there are still some unassembled regions which account ~10% of the whole genome missing in the first version (RS1) (Zhang et al., 2016a). By taking further efforts, we then improved both ZS97 and MH63 genome sequences to RS2 version which contained only several gaps in each assembly and immediately shared to public in 2018 (http://rice.hzau.edu.cn).

With high-coverage and accurate long-reads integrated with multiple assembling strategies in this study, we significantly improved our assemblies and successfully generated two gap-free genome assemblies of *xian* rice ZS97 and MH63, which are the first gap-free plant genome publicly available to date. Importantly, we had the first opportunity to study and compare the full centromeres of all chromosomes side by side in both rice varieties. More than one thousand genes were identified in rice centromere regions and ~16% of them were actively expressed. In addition, a graph-based rice pan-genome was built which contained presence/absence variations of 79 rice varieties. The two gap-free assemblies we present here will give scientists a clear picture of sequence divergence and how this impacts heterosis at the molecular level.

RESULTS

Generation and Annotation of ZS97 and MH63 Gap-free Genome Sequences

In this project, 56.73 Gb (~150X) and 86.85 Gb (~230X coverage) of PacBio reads (including both HiFi and CLR modes) were respectively generated for ZS97 and MH63 on PacBio Sequel II platform (Supplementary Figure 1; Supplementary Table 1). The PacBio HiFi and CLR reads were separately assembled with multiple *de novo* assemblers including Canu (Koren, Walenz et al. 2017), FALCON (Carvalho, Dupim et al. 2016), MECAT2 (Xiao, Chen et al. 2017) *etc*, and then the assembled contigs were merged through GPM pipeline (Zhang, Kudrna et al. 2016) (Supplementary Fig. 1; Supplementary Table 2-3). Finally, we built two gap-free reference genomes, named as ZS97RS3 and MH63RS3, which contained 12 pseudomolecules with total lengths of 391.56 Mb and 395.77 Mb, respectively (Fig. 1a; Table 1). Compared with the previous bacterial artificial chromosome (BAC) based genomes RS1, the RS3 assemblies gained 36-44 Mb additional sequences by filling all gaps in both ZS97RS1 and MH63RS1 (223 and 167 gaps, respectively) (Supplementary Table 4). Meanwhile, we corrected a few wrongly orientated or misassembled regions in RS1 sequences (e.g. the 6 Mb inversion on Chr06) (Supplementary Fig. 2a-c; Supplementary Table 4). The increased sequence mainly consisted of transposable elements and centromeres (Supplementary Fig. 2d). By detecting the 7-base telomeric repeat (CCCTAAA at 5' end or TTTAGGG at 3' end), we identified 19 and 22 telomeres that conducted 7 and 10 gapless telomere-to-telomere (T-to-T) pseudomolecules in ZS97RS3 and MH63RS3, respectively **(**Fig. 1a; Supplementary Table 5-6). In addition, the data obtained by different sequencing technologies have different coverage, both the PacBio HiFi and CLR reads covered >99.9% of the ZS97RS3 and MH63RS3 gap-free genomes, while BAC reads only covered 88.59% and 90.95%, respectively (Fig. 1b). The accuracy and completeness of the RS3 assemblies were further validated by 1) both Hi-C sequencing analysis and BioNano optical maps that showed high consistency with all pseudomolecules (Supplementary Fig. 3; Supplementary Table 2); 2) high mapping rates with various sequences, such as paired-end short reads from 48 RNA-seq libraries, paired BAC-end sequences, raw HiFi/CLR/Illumina reads from ZS97 and MH63 were obtained (Supplementary Table 7-9). 3) ZS97RS3 and MH63RS3 both captured 99.88% of the BUSCO reference gene set (Supplementary Table 10). 4) Long terminal repeat (LTR) annotation revealed the LTR assembly index (LAI) of ZS97RS3 and MH63RS3 were 24.01 and 22.74, respectively, which meet the standard of gold/platinum reference genomes (Ou et al., 2018; Mussurova et al. 2020) (Table 1). 5) More than twenty hundred thousand rRNAs were identified in ZS97RS3 and MH63RS3 (Supplementary Fig. 4), which were rarely identified in RS1. Furthermore, the evenly distributed breakpoints of aligned short and long reads indicated all sequence connections are of high accuracy at the single-base level in our final assemblies (Supplementary Fig. 5).

With the gap-free assemblies, we identified 465,242 transposable elements (TEs, ~181.00 Mb in total length) in ZS97RS3 and 468,675 TEs (~182.26Mb) in MH63RS3 (Supplementary Table 11-12), which accounted ~46.16% and ~45.99% of each genome and were higher than that in RS1 (~41.28% and ~41.58%). The increased portion mostly due to that an updated TE library and the closed gaps are primarily in TE-rich regions. TE contents in closed gap regions were 82.86% in ZS97RS3 and 84.17% in MH63RS3. We employed MAKER-P (Campbell et al., 2014) to annotate ZS97RS3 and MH63RS3 with all the same EST, RNA-Seq, and protein evidence as used in RS1 (Supplementary Fig. 1). In order to keep annotations consistent in different assembly versions, 51,027 gene models in ZS97RS1 and 50,341 in MH63RS1 were retained and migrated into RS3 version. Combining models annotated 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), which was 4,648 and 2,082 more than in RS1, respectively. More than 92% of annotated genes were supported by homologies with known proteins or functional domains in other species (Supplementary Table 13-14). The protein-coding non-TE genes were unevenly distributed across each chromosome with gene density increasing toward the chromosome ends (Supplementary Fig. 6). In addition, non-coding RNAs were annotated, including 636 and 618 transfer RNAs (tRNAs), 267,347 and 232,845 ribosomal RNAs (5S, 5.8S, 18S and 28S rRNAs), 582 and 586 small nucleolar RNAs (snRNAs), 1,550 and 1,568 microRNAs in ZS97RS3 and MH63RS3 (Supplementary Fig. 4).

There were 1.35 million single nucleotide polymorphisms (SNPs) and 0.26 million insertions/deletions (InDels) between ZS97 and MH63. This is relatively lower than the 2.56 million (2.58 million) SNPs and 0.48 million (0.49 million) InDels between ZS97 (MH63) and Nipponbare (Supplementary Fig. 6; Supplementary Table 15), confirming that intra-subspecies variations (*xian* vs. *xian*) were much less than inter-subspecies (*xian* vs. *geng*) variation. About 39% of non-TE genes (*i.e.* 15,526 models) in ZS97RS3 and MH63RS3 had syntenic position and highly identical sequences with synonymous SNPs. The remaining non-TE genes were categorized into four types: (1) 3,830 gene-pairs had the same length and syntenic positions, but contained nonsynonymous substitutions with identity ≥80%; (2) 10,886 gene-pairs were conserved with syntenic chromosomal locations, and protein sequences identity ≥ 80% and coverage > 50%; (3) 7,786 (ZS97RS3) and 7,704 (MH63RS3) non-TE genes were classified as “divergent genes”, which resulted from structural variations (SVs) between the two genomes; (4) 1,230 ZS97-specific genes and 1,460 MH63-specific genes were identified. The extensive gene structure difference between ZS97 and MH63 likely forms the basis of heterosis in their hybrids (Supplementary Table 16).

Location and Analyses of Centromeres in *Xian* Rice

Centromeres are essential for maintaining the integrity of the chromosome during cell division, and it ensures the fidelity of the chromosomes during inheritance.

Nevertheless, centromeres remain under-explored, especially in larger genomes (Perumal, Koh, et al. 2020). We identified the centromere regions of ZS97RS3 and MH63RS3 by ChIP-seq using rice CENH3 antibody (Fig. 2a-b). FISH analysis using ChIPed DNA revealed bright hybridization signal in the metaphase chromosomes indicating the presence of centromeric DNA sequences (Fig. 2b). Using MH63RS3 as the reference, for the first time, we determined that the lengths of rice centromeres are varied between 0.8-1.8 Mb (Supplementary Fig. 6-7; Supplementary Table 17-18). Rice centromeres consist of abundant repetitive sequences (78-80%), with representative LTR retrotransposons such as LTR/Gypsy (Supplementary Table 19-20). We classified rice centromeres into core and shell regions. Core centromere regions (CCRs) were identified by sequence homology to the 155-165 bp centromere-specific (*CentO*) satellite repeats (Cheng Z, et al. 2002), while shell regions were determined with the ChIP-seq signals. The length of CCRs ranged from 76 kb to 726 kb in different chromosomes with a total length 3.47 Mb in MH63RS3 (Supplementary Fig. 7, Supplementary Table 17). We manually checked the entire centromere regions (especially the boundary regions) of MH63RS3 and ZS97RS3 and found that the HiFi/CLR reads were evenly mapped with no ambiguous breakpoints (Fig. 2c, Supplementary Fig. 8), which evidences the high integrity and correctness of all assembled centromeres.

Comparative analysis showed that CCRs contain a few non-TE genes but a large amount of *CentO* satellite sequences (Fig. 2d; Supplementary Fig. 9). While shell regions contained >1,400 genes (~16% expressed), which include many centromere-specific retrotransposon sequences (Fig. 2d; Supplementary Table 21-23). For example, the Chr01 centromere of MH63RS3 is 1.6 Mb, and its CCR is ~726 kb containing 3,228 *CentO* sequences and 47 genes. The shell region on both sides of the CCR contained 114 *CentO* sequences and 61 none-TE genes (Fig. 2d; Supplementary Table 18; Supplementary Table 21). Only a very small number of genes located in the CCR can be transcribed and expressed, however, many genes in the shell regions are actively expressed (Fig. 2d). We also found that the methylation level of CG and CHG in the centromeric region was two-fold higher than that of the whole genome (Supplementary Table 24). This phenomenon is particularly prominent in *CentO* clustered regions.

Based on the complete centromere location, we counted the length and depth of the reads in both centromere and non-centromere regions. Although the centromere regions had slightly lower depth of reads than non-centromere regions (Supplementary Fig. 9b), which may be caused by highly repetitive elements. Overall, the average read length and coverage in centromere regions were broadly in line with non-centromere regions (Supplementary Fig. 9b). In addition, the proportion of LTR/gypsy accounting for over 90% of TEs in the centromere region is extremely higher than that of other types (Supplementary Fig. 9c), which is an obvious barrier to fully assembled.

To assess the conservation of rice centromeres, we identified centromeres and their core regions in 15 rice accessions with high-quality genomes (Zhou et al. 2020) (Supplementary Table 25). We observed that the lengths of CCRs in different chromosomes were significantly different, even for the same chromosome, the CCR lengths are also varied widely in different rice varieties (Supplementary Table 26). This reflected that the length and copy number of *CentO* repeats were not consistent in rice centromeres. For ZS97 and MH63, 72% conserved gene families were identified in centromere regions (Supplementary Fig. 9d). GO analysis showed that genes in ZS97 and MH63 centromere regions had similar functions (Supplementary Fig. 10b, c; Supplementary Table 27-28), which were significantly enriched in the GO term of ‘transcription from RNA polymerase III promoter’, ‘nucleic acid binding’ and ‘nucleoplasm part’, indicating the conservation of centromere function (Supplementary Fig. 10a**).** To better understand the long-range organization and evolution of the CCRs, we generated a heat map showing pairwise sequence identity of 1 kb along the centromeres (Supplementary Fig. 11a), and observed that the *CentO* sequences had the highest similarity in the middle and declined to both sides (Supplementary Fig. 11a). Furthermore, the profile of *CentO* sequences

(Supplementary Fig. 11b) illustrated the conservation of rice centromeres on the genomic level.

Graph-based Pan-genome and Pan-NLRome of Rice

Although several linear rice pan-genome had been constructed, the sequence was mainly based on *de novo* assembly of short-read re-sequencing data (Wang et al., 2018). In addition to 66 short-reads assembled genomes (Zhao et al., 2018), 13 genomes assembled by long-reads were selected to construct pan-genome (Supplementary Table 29). The above 79 rice varieties (7 *O. sativa aus*, 27 *indica/xian*, 25 *temperate japonica/geng*, 6 *tropical japonica/geng*, 1 *O. sativa aromatic* and 13 *O. rufipogon*) represent the major of *O. sativa* and *O. rufipogon* groups (Supplementary Table 29). Phylogenetic tree was constructed by using jacard similarity between long-kmer datasets to determine the similarity between different genomes. From the phylogenetic relationship, it was obvious that the same subgroups of Asian cultivated rice were clustered together, including *temperate japonica/geng, tropical japonica/geng*, *indica/xian* and *aus* (Fig 3a). It can also be observed that *xian* and *geng* were close to different subgroups of wild rice (Wing et al., 2018a; Xie et al., 2020). ZS97 and MH63 were in different branches in the *O. sativa xian* subgroup (Fig 3a). Previous studies had divided them into the *indica/xian II* and *indica/xian I* subgroups respectively, which represented different *O. sativa indica* population and showed a large genetic difference (Xie et al., 2015). We used the gap-free genome MH63RS3 as the reference to identify presence/absence variations (PAVs) in other rice varieties to construct graph-based pan-genome of *O. sativa* (Os-GPG), which can not only identify complex SVs, but also improve the accuracy of variation calls around SVs (Rakocevic et al., 2019; Liu et al., 2020). After filtering redundancy and decontamination, the pan-PAVs of Asian cultivated rice is ~320 Mb, of which *xian* is 169 Mb and *geng* is 145 Mb (Fig. 3a; Supplementary Table 30). Affected by the diversity of *xian* rice, the PAV of *xian* rice was greater, even when the *xian* genome was used as the reference. 17,365 protein-coding genes were annotated in pan-PAVs of Asian cultivated rice that were not present in reference genome (Supplementary Table 30). We merged 454,187 PAVs from all genomes into a set of 278,567 nonredundant PAVs. Further, vg toolkit (Garrison et al., 2018) was used to construct a graph-based pan-genome of rice, which can be directly used for read mapping and GWAS analysis (Fig. 3b). It is the first graph-based pan-genome obtained from a gap-free reference genome in rice. The pan-PAVs sequence had a lower gene density than reference, but contained abundant resistance genes (NLRs). We identified 557 NLRs in pan-PAVs, and this number is similar to the reference genomes (MH63:509; Nip: 473 (Wang et al., 2019)) (Supplementary Table 31). Therefore, when a single reference genome was used to study the adaptability of rice, almost half of the NLR genes are missed. A large number of NLRs were imbalanced in *'xian'* and *'geng'* subgroups, and some NLRs only existed in a few wild rice varieties (Supplementary Fig. 12b). The Os-PGP provides valuable resources and should promote rice studies in the post-genomic era. The distribution of PAVs and NLRs of ZS97 and MH63 were similar in other chromosomes, while highly different in the end of chromosome 11 (Fig. 3c, Supplementary Fig. 12a). In this region, we found two large SVs, named MH-Ex1 and MH-INS1, between ZS97 and MH63 (Supplementary Fig. 13a). Through mapping the PacBio HiFi reads of ZS97 and MH63 to the end of chromosome 11 of MH63RS3 genome, we clearly observed the two large SVs. The reads of MH63 can continuously span these two regions, while ZS97 reads cannot cover these regions (Fig. S11b). For MH-Ex1, most of the resistance genes in ZS97 amplified 2-10 times in MH63 (Fig. 3d; Supplementary Table 32), resulting a large genomic sequence expansion (from 0.18 Mb in ZS97 to 0.82 Mb in MH63). It is very interesting that most of the expanded resistance genes are not expressed or lowly expressed in most tissues except root (Fig. 3d; Supplementary Fig. 13c; Supplementary Table 32). For MH-INS1, MH63RS3 genome had an 857 kb insertion compared with ZS97RS3 genome, including eleven resistance genes with low expression levels in most tissues except root (Supplementary Table 33). We further scanned the two SVs (MH-Ex1 and MH-INS1) in the remaining 25 rice genomes assembled based on PacBio long-read sequencing, and observed that MH-Ex1 and MH-INS1 were incomplete in all the other rice varieties compared with MH63 genome (Zhou et al. 2020) (Fig. 3d, Supplementary Fig. 14; Supplementary Table 34). The above example indicated the genetic advantage of MH63 a donor of resistance genes. This was an illustration that Os-PGP will provide the full range of short to long-range SVs that exist across the *O. sativa*.

In summary, we assembled two gap-free genomes of *xian* rice ZS97 and MH63, which are the first report of gapless plant genomes up to now. Based on these genomes, we analyzed and compared the complete centromeres of all chromosomes in both rice varieties, and observed that >1,500 genes were existed in centromere regions and ~16% of them were actively expressed. Based on the gap-free MH63RS3 genome, a graph-based rice pan-reference-genome was constructed containing presence/absence variations of 79 rice varieties, which can be used as a solid foundation for further genome wide association studies.

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 previous study (Pendleton, M. et al. 2015). The genomes of MH63 and ZS97 were sequenced using PacBio Sequel II platform (Pacific Biosciences), including 8.34 Gb HiFi reads (~23x coverage) and 48.39Gb CLR reads (~131x coverage) for ZS97, and 37.88 Gb HiFi reads (~103x coverage) and 48.97 Gb CLR reads (~132x coverage) for MH63 genomes. Plant tissues were extracted using the BioNano plant tissue extraction protocol. We embedded the extracted DNA in BioRad LE agarose for subsequent washes of TE, proteinase K (0.8mg/ml), and RNAse A (20μL/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 (0.1μm) with TE to equilibrate ion concentrations. We then nicked the DNA with a nickase restriction enzyme BssSI (2U/μL) with a 6 bp sequence recognition motif. Labeled nucleotides were incorporated at breakpoints and the DNA was counterstained. Each sample was loaded onto 2 nanochannel flow cells of an Irys machine for DNA imaging. Truseq Nano DNA HT Sample preparation Kit following manufacturer's standard protocol (Illumina) was used to generate the libraries for Illumina paired-end genome sequencing. These libraries were sequenced to generate 150 bp paired-end reads by Illumina HiSeq X Ten platform with 350 bp insert size.

Genome Assembly and Assessment

In this work, seven tools based on different algorithms were performed to assemble the genomes of ZS97 and MH63. (1) Canu v1.8 (Koren S et al., 2017) was used to assemble the genomes with default parameters; (2) FALCON toolkit v0.30 (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 -M24 -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 2.6-release (Kolmogorov et al., 2019) was set with “--genome-size 400m”; (5) Wtdbg2 2.5 (Ruan et al., 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 polish and correct the wtdbg2 assembly results; (6) NextDenovo v2.1-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 (Heng Li 2016) with default parameters. Based on these seven software, Genome Puzzle Master GPM (Zhang et al., 2016) was performed to integrate and optimize the assembled contigs, and visualize the complete chromosomes. Based on the HiFi and CLR sequencing data, we used GenomicConsensus package of SMRTLink/7.0.1.66975 (https://www.pacb.com/support/) to polish the assembled genome twice with Arrow algorithm, the parameters are: --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 repeated twice. Molecules were then assembled using 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 was visualized using IrysView v2.5.1. These assemblies were used with draft genome assemblies to validate and scaffold the sequences. Bionano map data was aligned to the merged contigs using RefAlignerAssembler in IrysView software to do the verification.

ZS97RS3 and MH63RS3 genome completeness assessment using BUSCO v4.0.6 (Felipe A et al., 2015). Besides, we mapped the PacBio HiFi reads and PacBio CLR reads (using Minimap2 (Li 2018)), Illimina reads (using BWA-0.7.17 (Jo H et al., 2015)), BES/BAC reads (using BLASTN v2.7.1 (Altschul et al., 1990)), Hi-C reads (using HiC-Pro v2.11.1 (Servant et al., 2015)), RNA-Seq reads (using Hisat2 v2.1.0 (Kim et al., 2015)) to both genome assemblies find both assemblies performed well.

Gene and Repeat Annotations

MAKER-P (Campbell et al., 2014) version 3 was used to annotate the ZS97RS3 and MH63RS3 genomes. All the evidences are the same as that used for RS1 genomes. To ensure the consistency of RS1 version, genes can completely map to RS3 genome were retained. The new genes in gap regions were obtained from MAKER-P (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 E-value 1e-10. tRNAs were identified with tRNAscan-SE (Lowe, T. M. & Eddy, S. R. 19997) using default parameters; rRNA genes were identified by searching the genome assembly against the rRNA sequences of Nipponbare using BLASTN v2.7.1 (Altschul et al., 1990); miRNAs and snRNAs were predicted using INFERNAL of Rfam (Griffiths-Jones, S. et al., 2005) (v14.1). Repeats in the genome were annotated using RepeatMasker (Smit et al. 2015) with RepBase (Bao et al., 2015), TIGR Oryza Repeats (v3.3) with RMBlast search engine. For the overlapping repeats in different classes, LTR retrotransposons were kept first, next TIR, and then SINE and LINE, finally helitrons. This priority order was based on stronger structural signatures. Besides, the known nested insertions 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 Z, Wang H 2007) was used to identify complete LTR-RTs with target site duplications (TSDs), primer binding sites (PBS) and polypurine tract (PPT).

Chromatin Immunoprecipitation (ChIP) and ChIP-seq

The procedures for chromatin immunoprecipitation (ChIP) were adopted from Nagaki et al. (2003) and Walkowiak et al. (2020). The nuclei were isolated from 4-week-old seedlings. The nuclei were digested with micrococcal nuclease (Sigma-Aldrich, St.

Louis, MO) to liberate nucleosomes. The digested mixture was incubated overnight with 3 µg of rice CENH3 antibody at 4°C. The target antibodies were captured from the mixture using Dynabeads Protein G (Invitrogen, Carlsbad, CA). ChIP-seq libraries were constructed using TruSeq ChIP Library Preparation Kit (Illumina, San Diego, CA) following the manufacturer's instructions and the libraries were sequenced on Illumina HiSeqX10 with 2x150 bp sequencing run.

Fluorescence *in situ* Hybridization (FISH)

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 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield antifade solution (Vector Laboratories, Burlingame, CA). The images were captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy LLC, 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 genomes, we use BLAST (Altschul et al., 1990) to align the c *CentO* satellite repeats in rice to the reference genome with E-value 1e-5, then use

BEDtools (Quinlan et al., 2014) to merge the result with the parameter -d 50000. Then, from the outside to the inside, if the number of consecutive *CentO* is less than 5, it is classified as core region if the number of consecutive *CentO* is greater than 5 but less than 10, and the distance between two *CentO* clusters a less than 10kb, it is classified as core region; if the number of consecutive *CentO* is more than 10, it is directly classified as core region.

For the identification of the whole centromere region, we use BWA-0.7.17 (Jo H et al., 2015) to align the CENH3 ChIP-Seq reads to MH63RS3 and ZS97RS3 genomes, and use SAMtools (Li H et al., 2009) to filter the results with mapQ value above 30; then we use MACS2 (Zhang Y et al., 2008) to call the peaks of CENH3. Finally, we combined the distribution of CENH3 histone, CentOS, repeats and genes to jointly define all the centromeric region of MH63RS3 and ZS97RS3 genomes. It should be noted that when determining the peaks of CENH3 histones, the standard is that if three consecutive peaks value > 30 and no cluster interference, the last peak position is defined as the centromeric boundary position; if three consecutive peaks value > 30 but there has cluster interference, reduce the peak value standard to 20, and then define the centromere boundary; combined with manual adjustment of the position.

To compare of *CentO* sequence similarity, first we use BEDtools (Quinlan et al., 2014) to obtain sequences of centromere core regions, and divide them into 1 kb continuous sequences; then we use Minimap2 (Li 2018) to align the sequences, the parameters are: -f 0.00001 -t 8 -X --eqx -ax ava -pb; finally, we use a custom python script to filter the result file, and use R to generate a heat map showing pairwise sequence identity (Logsdon, Vollger et al. 2020).

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 WJ 2002) to search telomere-associated tandem repeats sequence (TAS) from TIGR Oryza Repeat database (Ouyang et al., 2004) in whole genome.

Identification of PAVs

We selected 79 rice varieties to construct phylogenetic tree, 66 were from previous studies (Zhao et al., 2018) and 11 were downloaded from NCBI (as of 1-30-2020). Sourmash was used to compute hash sketches from genome sequences (k-mer = 301) and calculate jaccard similarity of 79 rice genomes to generate phylogenetic tree (Pierce et al., 2019). The rice genomes were aligned to reference genome MH63 using Mummer(4.0.0beta2) (Marçais et al., 2018) with parameters settings '-c 90 -l 40'. Then used "show-diff" to select for unaligned regions. Further we merged all *O. sativa indica* and *O. sativa japonica* unaligned sequences and then used CD-HIT(v4.8.1) (Fu et al., 2012) to remove redundant sequences. Finally, we used blastn to remove contaminate sequences with parameters settings '-evalue 1e-5 -best_hit_overhang 0.25 -perc_identity 0.5 -max_target_seqs 10' and the rest is PAVs sequences.

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. NLRs were defined to contain at least NB, a TIR, or a CCR(RPW8) domain and we classified NLRs based on above structural features. NLRs domain contain only 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 Orthologues

MCscan (python version) (Tang et al., 2008) was used to identify collinear orthologues between chromosome 11 of ZS97RS3 and MH63RS3 genomes with default parameters.

Construction of Graph-based Pan-genome

MH63RS3 was set as a reference and the pan-PAVs sequences were saved in variant call format (VCF). The graph-based pan-genome was construct via the vg (https://github.com/vgteam/vg, version v1.29.0) toolkit (Garrison et al., 2018) with default parameters.

DATA AVAILABILITY

All the raw sequencing data generated for this project are achieved at NCBI under accession numbers SRR13280200, SRR13280199 and SRR13288213 for ZS97, SRX6957825, SRX6908794, SRX6716809 and SRR13285939 for MH63. The genome assemblies are available at NCBI (CP056052-CP056064 for ZS97RS3, CP054676-CP054688 for MH63RS3) and annotations are visualized with Gbrowse at http://rice.hzau.edu.cn. All the materials in this study including introgression lines are available upon request.

FUNDING

This research was supported by the National Key Research and Development Program of China (2016YFD0100904), the National Natural Science Foundation of China (31871269), Hubei Provincial Natural Science Foundation of China (2019CFA014), the 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 FISH 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., Y.H. J.-W.F., W.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.

ACKNOWLEDGEMENTS

We sincerely thank 1) Pacific Biosciences of California, Inc. for sequencing of MH63, 2) Wuhan Frasergen Bioinformatics Co., Ltd. for sequencing of ZS97 and 3) Dr. Jiming Jiang at MSU for his critical comments and constructive suggestions on our centromere analyses.

ONLINE CONTENT

Any methods, additional references, Research reporting summaries, source data, statements of code and data availability and associated accession codes are available online.

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.

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.

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.

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.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics **28:**3150–3152.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T.,

Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. **36:**875–879.

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. USA **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.

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.

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.

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., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754–1760.

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

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell **182:**162–176.

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. bioRxiv 2020.09.08.285395

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.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature **585:**79–84.

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., 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.

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

Pierce, N.T., Irber, L., Reiter, T., Brooks, P., and Brown, C.T. (2019). Large-scale sequence comparisons with *sourmash*. F1000Res **8:**1006.

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

Rakocevic, G., Semenyuk, V., Lee, W.P., Spencer, J., Browning, J., Johnson, I.J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M.C., et al. (2019). Fast and accurate genomic analyses using genome graphs. Nat. Genet. **51:**354–362.

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.

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.

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.

Wang, L., Zhao, L., Zhang, X., Zhang, Q., Jia, Y., Wang, G., Li, S., Tian, D., Li, W.H., and Yang, S. (2019). Large-scale identification and functional analysis of NLR genes in blast resistance in the Tetep rice genome sequence. Proc. Natl. Acad. Sci. USA **116:**18479–18487.

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.

Wing, R.A., Purugganan, M.D., and Zhang, Q. (2018). The rice genome revolution: from an ancient grain to green super rice. Nat. Rev. Genet. **19:**505–517.

Xie, W., Wang, G., Yuan, M., Yao, W., Lyu, K., Zhao, H., Yang, M., Li, P., Zhang, X., Yuan, J., et al. (2015). Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc. Natl. Acad. Sci. USA **112:**E5411–5419.

Xie, X., Du, H., Tang, H., Tang, J., Tan, X., Liu, W., Li, T., Lin, Z., Liang, C., and Liu, Y.G. (2020). A chromosome-level genome assembly of the wild rice *Oryza rufipogon* facilitates tracing the origins of Asian cultivated rice. Sci China Life Sci. doi: 10.1007/s11427-020-1738-x.

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:**E5163–5171.

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.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. **50:**278–284.

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.

FIGURE LEGENDS

Fig. 1 | Two gap-free genomes of rice.

a). Collinearity analysis between ZS97RS3 and MH63RS3. The collinear regions between ZS97RS3 and MH63RS3 were linked as the gray lines. All the RS1 gap regions were closed in RS3 and showed in the yellow block. The black triangle indicated the telomere, there are 7 T-to-T chromosomes in ZS97RS3 (Chr01, Chr02, Chr03, Chr04, Chr06, Chr07, Chr11) and 10 T-to-T chromosomes in MH63RS3 (Chr01, Chr02, Chr03, Chr04, Chr05, Chr06, Chr07, Chr09, Chr10, Chr12). All the centromeres are complete and repeat length distribution diagrams were plotted above/under each chromosome; b). Histogram showed the reads coverage for different libraries in MH63RS3 and ZS97RS3, including BAC, CCS and CLR reads.

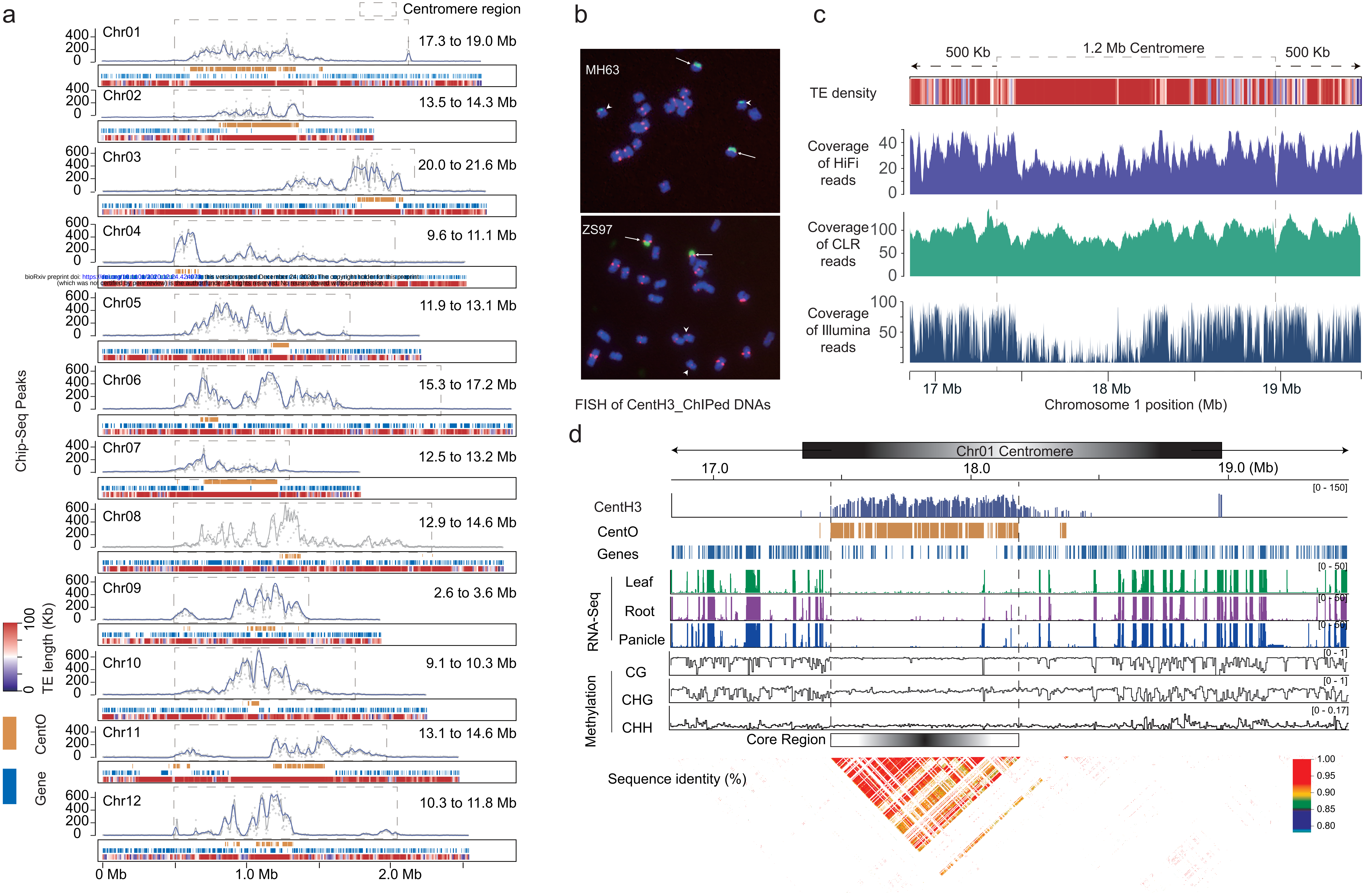
Fig. 2 | Complete rice centromeres.

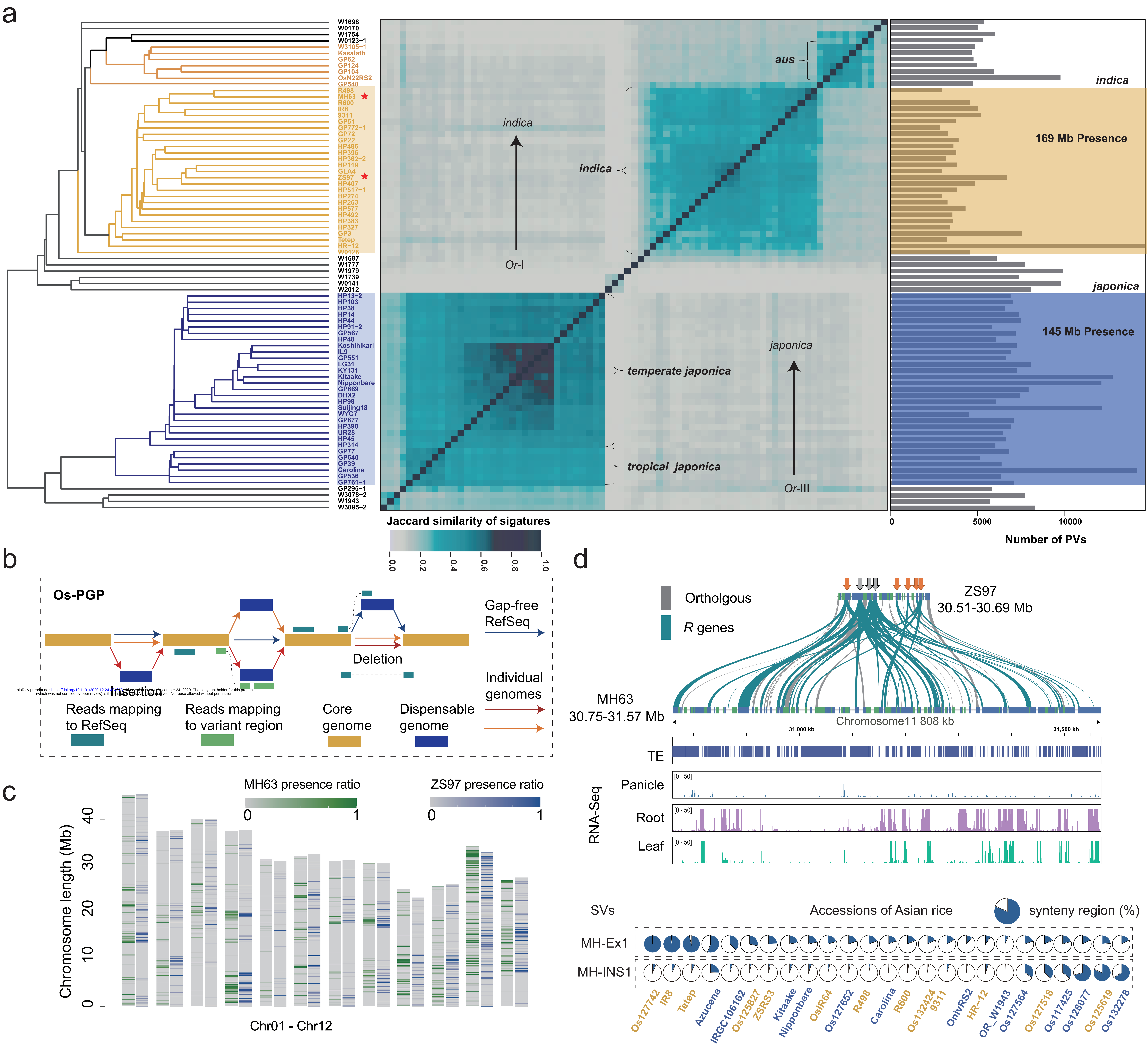
a, The definition of MH63RS3 centromere. the first to fourth layers indicate the histone CENH3 Chip-seq distribution, the *CentO* satellite distribution, t genes distribution, and of TE distribution, respectively. The dotted frame represents the final centromere area. **b**, FISH signals detected in metaphase of meiosis for MH63RS3 and ZS97RS3, white arrows indicate DNA elements in the centromeric region. **c**, Coverage of HiFi, CLR, Illumina reads and distribution of TEs in the centromere on Chr01 (extended 500 kb left and right) of MH63RS3. **d**, Characteristics of the centromere on Chr01 of MH63RS3. The first layer is histone CENH3 distribution, the second layer is the CentOS distribution, the third layer is the Genes distribution, the fourth to sixth levels are gene expression, the seventh to ninth levels are methylation distribution, the tenth layer is CentOS sequence similarity.

Fig. 3 | The graph-based pan-genome and pan-NLRome of rice.

Figure 3. a, Phylogenetic tree of the 79 rice varieties. 79 rice varieties phylogenetic tree (left), black represents wild rice varieties, orange represents *O. sativa aus*, Orange shadow represents *O. sativa indica*, blue shadow represents *O. sativa japonica*, heat map represents the jaccard similarity of pairwise rice (middle), and bar graph represents the number of PV per rice (right). **b,** The schematic diagram of rice graph-based pan-genome. **c,** Distribution of the difference regions between ZS97RS3 and MH63RS3 on the chromosome. **d,** The expansion structural variation of MH63RS3. The expansion structural variation at the end of chromosome 11 of MH63RS3, from top to bottom are the gene collinearity of ZS97RS3 and MH63RS3, the TE distribution, the gene expression in this region and coverage ratio of two structural variations in 25 rice varieties.

a
0Mb
10Mb
20Mb
30Mb
40Mb
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
ZS97RS3
MH63RS3
Closed
Telomere
Centromere
100
0
100
Repeat Length (%)
b
Reads Coverage (%)
100
75
50
25
0
90.95
99.99
99.99
88.59
99.96
99.98
MH63
ZS97
Accessions
Type
BAC
CCS
CLR


a
Centromere region
Chr01 17.3 to 19.0 Mb
Chr02 13.5 to 14.3 Mb
Chr03 20.0 to 21.6 Mb
Chr04 9.6 to 11.1 Mb
Chr05 11.9 to 13.1 Mb
Chr06 15.3 to 17.2 Mb
Chr07 12.5 to 13.2 Mb
Chr08 12.9 to 14.6 Mb
Chr09 2.6 to 3.6 Mb
Chr10 9.1 to 10.3 Mb
Chr11 13.1 to 14.6 Mb
Chr12 10.3 to 11.8 Mb
0 200 400 600
Chip-Seq Peaks
TE length (Kb)
100
0
CentO
Gene
0 Mb
1.0 Mb
2.0 Mb
b
MH63
ZS97
FISH of CentH3_ChIPed DNAs
c
500 Kb
1.2 Mb Centromere
500 Kb
TE density
Coverage of HiFi reads
40 20 0
Coverage of CLR reads
100 50 0
Coverage of Illumina reads
100 50
17 Mb
18 Mb
19 Mb
Chromosome 1 position (Mb)
d
Chr01 Centromere
17.0
18.0
19.0 (Mb)
CentH3
[0 - 150]
CentO
Genes
RNA-Seq
Leaf
Root
Panicle
[0 - 50]
Methylation
CG
CHG
CHH
[0 - 1]
[0 - 0.17]
Core Region
Sequence identity (%)
1.00
0.95
0.90
0.85
0.80


a
W1698
W0170
W1754
W0123−1
W3105−1
Kasalath
GP62
GP124
GP104
OsN22RS2
GP540
R498
MH63 ★
R600
IR8
9311
GP51
GP772−1
GP72
GP22
HP486
HP396
HP362−2
HP119
GLA4
ZS97 ★
HP407
HP517−1
HP274
HP263
HP577
HP492
HP383
HP327
GP3
Tetep
HR−12
W0128
W1687
W1777
W1979
W1739
W0141
W2012
HP13−2
HP103
HP38
HP14
HP44
HP91−2
GP567
HP48
Koshihikari
IL9
GP551
LG31
KY131
Kitaake
Nipponbare
GP669
DHX2
HP98
Suijing18
WYG7
GP677
HP390
UR28
HP45
HP314
GP77
GP640
GP39
Carolina
GP536
GP761−1
GP295−1
W3078−2
W1943
W3095−2
aus
indica
indica
Or-I
temperate japonica
tropical japonica
japonica
Or-III
Jaccard similarity of sigatures
0.0
0.2
0.4
0.6
0.8
1.0
indica
169 Mb Presence
japonica
145 Mb Presence
0
5000
10000
Number of PVs
b
Os-PGP
Gap-free RefSeq
Insertion
Deletion
Individual genomes
Reads mapping to RefSeq
Reads mapping to variant region
Core genome
Dispensable genome
c
MH63 presence ratio
0
1
ZS97 presence ratio
0
1
Chromosome length (Mb)
0
10
20
30
40
Chr01 - Chr12
d
Ortholgous
R genes
ZS97
30.51-30.69 Mb
MH63
30.75-31.57 Mb
Chromosome11 808 kb
31,000 kb
31,500 kb
TE
RNA-Seq
Panicle
Root
Leaf
[0 - 50]
[0 - 50]
[0 - 50]
SVs
Accessions of Asian rice
synteny region (%)
MH-Ex1
MH-INS1
Os127742
IR8
Tetep
Azucena
IRGC106162
Os125827
ZSRS3
Kitaake
Nipponbare
OsIR64
Os127652
R498
Carolina
R600
Os132424
9311
OnivRS2
HR-12
OR_W1943
Os127564
Os127518
Os117425
Os128077
Os125619
Os132278