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)

30

31 **ABSTRACT**

32 **Asia rice (*Oryza sativa*) is divided into two subgroups, *indica/xian* and**

33 ***japonica/geng*, the former has greater intraspecific diversity than the latter. Here,**

34 **for the first time, we report the assemblies and analyses of two gap-free *xian* rice**

35 **varieties 'Zhenshan 97 (ZS97)' and 'Minghui 63 (MH63)'. Genomic sequences of**

36 **these elite hybrid parents express extensive difference as the foundation for**

37 **studying heterosis. Furthermore, the gap-free rice genomes provide global**

38 **insights to investigate the structure and function of centromeres in different**

39 **chromosomes. All the rice centromeric regions share conserved**

40 **centromere-specific satellite motifs but with different copy numbers and**

41 **structures. Importantly, we show that there are >1,500 genes in centromere**

42 **regions and ~16% of them are actively expressed. Based on MH63 gap-free**

43 **reference genome, a graph-based rice pan-genome (Os-GPG) was constructed**

44 **containing presence/absence variations of 79 rice varieties. Compared with the**

45 **other rice varieties, MH63 contained the largest number of resistance genes. The**

46 **acquisition of ZS97 and MH63 gap-free genomes and graph-based pan-genome**

47 **of rice lays a solid foundation for the study of genome structure and function in**

48 **plants.**

49

50 **Key words:** gap-free genomes, ZS97, MH63, centromere structure, graph-based

51 pan-genome

52

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

3

82   clear picture of sequence divergence and how this impacts heterosis at the molecular

83   level.

84

85   **RESULTS**

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

87   In this project, 56.73 Gb (~150X) and 86.85 Gb (~230X coverage) of PacBio reads

88   (including both HiFi and CLR modes) were respectively generated for ZS97 and

89   MH63 on PacBio Sequel II platform (Supplementary Figure 1; Supplementary Table

90   1). The PacBio HiFi and CLR reads were separately assembled with multiple *de novo*

91   assemblers including Canu (Koren, Walenz et al. 2017), FALCON (Carvalho, Dupim

92   et al. 2016), MECAT2 (Xiao, Chen et al. 2017) *etc*, and then the assembled contigs

93   were merged through GPM pipeline (Zhang, Kudrna et al. 2016) (Supplementary Fig.

94   1; Supplementary Table 2-3). Finally, we built two gap-free reference genomes,

95   named as ZS97RS3 and MH63RS3, which contained 12 pseudomolecules with total

96   lengths of 391.56 Mb and 395.77 Mb, respectively (Fig. 1a; Table 1). Compared with

97   the previous bacterial artificial chromosome (BAC) based genomes RS1, the RS3

98   assemblies gained 36-44 Mb additional sequences by filling all gaps in both ZS97RS1

99   and MH63RS1 (223 and 167 gaps, respectively) (Supplementary Table 4). Meanwhile,

100   we corrected a few wrongly orientated or misassembled regions in RS1 sequences

101   (e.g. the 6 Mb inversion on Chr06) (Supplementary Fig. 2a-c; Supplementary Table 4).

102   The increased sequence mainly consisted of transposable elements and centromeres

103   (Supplementary Fig. 2d). By detecting the 7-base telomeric repeat (CCCTAAA at 5'

104   end or TTTAGGG at 3' end), we identified 19 and 22 telomeres that conducted 7 and

105   10 gapless telomere-to-telomere (T-to-T) pseudomolecules in ZS97RS3 and

106   MH63RS3, respectively (Fig. 1a; Supplementary Table 5-6). In addition, the data

107   obtained by different sequencing technologies have different coverage, both the

108   PacBio HiFi and CLR reads covered >99.9% of the ZS97RS3 and MH63RS3 gap-free

109   genomes, while BAC reads only covered 88.59% and 90.95%, respectively (Fig. 1b).

110   The accuracy and completeness of the RS3 assemblies were further validated by 1)

111 both Hi-C sequencing analysis and BioNano optical maps that showed high

112 consistency with all pseudomolecules (Supplementary Fig. 3; Supplementary Table 2);

113 2) high mapping rates with various sequences, such as paired-end short reads from 48

114 RNA-seq libraries, paired BAC-end sequences, raw HiFi/CLR/Illumina reads from

115 ZS97 and MH63 were obtained (Supplementary Table 7-9). 3) ZS97RS3 and

116 MH63RS3 both captured 99.88% of the BUSCO reference gene set (Supplementary

117 Table 10). 4) Long terminal repeat (LTR) annotation revealed the LTR assembly index

118 (LAI) of ZS97RS3 and MH63RS3 were 24.01 and 22.74, respectively, which meet the

119 standard of gold/platinum reference genomes (Ou et al., 2018; Mussurova et al. 2020)

120 (Table 1). 5) More than twenty hundred thousand rRNAs were identified in ZS97RS3

121 and MH63RS3 (Supplementary Fig. 4), which were rarely identified in RS1.

122 Furthermore, the evenly distributed breakpoints of aligned short and long reads

123 indicated all sequence connections are of high accuracy at the single-base level in our

124 final assemblies (Supplementary Fig. 5).

125   With the gap-free assemblies, we identified 465,242 transposable elements (TEs,

126 ~181.00 Mb in total length) in ZS97RS3 and 468,675 TEs (~182.26Mb) in MH63RS3

127 (Supplementary Table 11-12), which accounted ~46.16% and ~45.99% of each

128 genome and were higher than that in RS1 (~41.28% and ~41.58%). The increased

129 portion mostly due to that an updated TE library and the closed gaps are primarily in

130 TE-rich regions. TE contents in closed gap regions were 82.86% in ZS97RS3 and

131 84.17% in MH63RS3. We employed MAKER-P (Campbell et al., 2014) to annotate

132 ZS97RS3 and MH63RS3 with all the same EST, RNA-Seq, and protein evidence as

133 used in RS1 (Supplementary Fig. 1). In order to keep annotations consistent in

134 different assembly versions, 51,027 gene models in ZS97RS1 and 50,341 in

135 MH63RS1 were retained and migrated into RS3 version. Combining models

136 annotated with MAKER-P in the newly assembled regions, the final annotations in

137 ZS97RS3 and MH63RS3 contained 60,935 and 59,903 gene models, of which 39,258

138 and 39,406 were classified as non-TE gene loci (Table 1), which was 4,648 and 2,082

139 more than in RS1, respectively. More than 92% of annotated genes were supported by

140 homologies with known proteins or functional domains in other species
141 (Supplementary Table 13-14). The protein-coding non-TE genes were unevenly
142 distributed across each chromosome with gene density increasing toward the
143 chromosome ends (Supplementary Fig. 6). In addition, non-coding RNAs were
144 annotated, including 636 and 618 transfer RNAs (tRNAs), 267,347 and 232,845
145 ribosomal RNAs (5S, 5.8S, 18S and 28S rRNAs), 582 and 586 small nucleolar RNAs
146 (snRNAs), 1,550 and 1,568 microRNAs in ZS97RS3 and MH63RS3 (Supplementary
147 Fig. 4).

148 There were 1.35 million single nucleotide polymorphisms (SNPs) and 0.26 million
149 insertions/deletions (InDels) between ZS97 and MH63. This is relatively lower than
150 the 2.56 million (2.58 million) SNPs and 0.48 million (0.49 million) InDels between
151 ZS97 (MH63) and Nipponbare (Supplementary Fig. 6; Supplementary Table 15),
152 confirming that intra-subspecies variations (*xian* vs. *xian*) were much less than
153 inter-subspecies (*xian* vs. *geng*) variation. About 39% of non-TE genes (*i.e.* 15,526
154 models) in ZS97RS3 and MH63RS3 had syntenic position and highly identical
155 sequences with synonymous SNPs. The remaining non-TE genes were categorized
156 into four types: (1) 3,830 gene-pairs had the same length and syntenic positions, but
157 contained nonsynonymous substitutions with identity ≥80%; (2) 10,886 gene-pairs
158 were conserved with syntenic chromosomal locations, and protein sequences identity
159 ≥ 80% and coverage > 50%; (3) 7,786 (ZS97RS3) and 7,704 (MH63RS3) non-TE
160 genes were classified as "divergent genes", which resulted from structural variations
161 (SVs) between the two genomes; (4) 1,230 ZS97-specific genes and 1,460
162 MH63-specific genes were identified. The extensive gene structure difference
163 between ZS97 and MH63 likely forms the basis of heterosis in their hybrids
164 (Supplementary Table 16).

165

166 **Location and Analyses of Centromeres in *Xian* Rice**
167 Centromeres are essential for maintaining the integrity of the chromosome during cell
168 division, and it ensures the fidelity of the chromosomes during inheritance.

169 Nevertheless, centromeres remain under-explored, especially in larger genomes

170 (Perumal, Koh, et al. 2020). We identified the centromere regions of ZS97RS3 and

171 MH63RS3 by ChIP-seq using rice CENH3 antibody (Fig. 2a-b). FISH analysis using

172 ChIPed DNA revealed bright hybridization signal in the metaphase chromosomes

173 indicating the presence of centromeric DNA sequences (Fig. 2b). Using MH63RS3 as

174 the reference, for the first time, we determined that the lengths of rice centromeres are

175 varied between 0.8-1.8 Mb (Supplementary Fig. 6-7; Supplementary Table 17-18).

176 Rice centromeres consist of abundant repetitive sequences (78-80%), with

177 representative LTR retrotransposons such as LTR/Gypsy (Supplementary Table

178 19-20). We classified rice centromeres into core and shell regions. Core centromere

179 regions (CCRs) were identified by sequence homology to the 155-165 bp

180 centromere-specific (*CentO*) satellite repeats (Cheng Z, et al. 2002), while shell

181 regions were determined with the ChIP-seq signals. The length of CCRs ranged from

182 76 kb to 726 kb in different chromosomes with a total length 3.47 Mb in MH63RS3

183 (Supplementary Fig. 7, Supplementary Table 17). We manually checked the entire

184 centromere regions (especially the boundary regions) of MH63RS3 and ZS97RS3 and

185 found that the HiFi/CLR reads were evenly mapped with no ambiguous breakpoints

186 (Fig. 2c, Supplementary Fig. 8), which evidences the high integrity and correctness of

187 all assembled centromeres.

188 Comparative analysis showed that CCRs contain a few non-TE genes but a large

189 amount of *CentO* satellite sequences (Fig. 2d; Supplementary Fig. 9). While shell

190 regions contained >1,400 genes (~16% expressed), which include many

191 centromere-specific retrotransposon sequences (Fig. 2d; Supplementary Table 21-23).

192 For example, the Chr01 centromere of MH63RS3 is 1.6 Mb, and its CCR is ~726 kb

193 containing 3,228 *CentO* sequences and 47 genes. The shell region on both sides of the

194 CCR contained 114 *CentO* sequences and 61 none-TE genes (Fig. 2d; Supplementary

195 Table 18; Supplementary Table 21). Only a very small number of genes located in the

196 CCR can be transcribed and expressed, however, many genes in the shell regions are

197 actively expressed (Fig. 2d). We also found that the methylation level of CG and CHG

7

198   in the centromeric region was two-fold higher than that of the whole genome

199   (Supplementary Table 24). This phenomenon is particularly prominent in *CentO*

200   clustered regions.

201   Based on the complete centromere location, we counted the length and depth of

202   the reads in both centromere and non-centromere regions. Although the centromere

203   regions had slightly lower depth of reads than non-centromere regions

204   (Supplementary Fig. 9b), which may be caused by highly repetitive elements. Overall,

205   the average read length and coverage in centromere regions were broadly in line with

206   non-centromere regions (Supplementary Fig. 9b). In addition, the proportion of

207   LTR/gypsy accounting for over 90% of TEs in the centromere region is extremely

208   higher than that of other types (Supplementary Fig. 9c), which is an obvious barrier to

209   fully assembled.

210   To assess the conservation of rice centromeres, we identified centromeres and

211   their core regions in 15 rice accessions with high-quality genomes (Zhou et al. 2020)

212   (Supplementary Table 25). We observed that the lengths of CCRs in different

213   chromosomes were significantly different, even for the same chromosome, the CCR

214   lengths are also varied widely in different rice varieties (Supplementary Table 26).

215   This reflected that the length and copy number of *CentO* repeats were not consistent

216   in rice centromeres. For ZS97 and MH63, 72% conserved gene families were

217   identified in centromere regions (Supplementary Fig. 9d). GO analysis showed that

218   genes in ZS97 and MH63 centromere regions had similar functions (Supplementary

219   Fig. 10b, c; Supplementary Table 27-28), which were significantly enriched in the GO

220   term of 'transcription from RNA polymerase III promoter', 'nucleic acid binding' and

221   'nucleoplasm part', indicating the conservation of centromere function

222   (Supplementary Fig. 10a). To better understand the long-range organization and

223   evolution of the CCRs, we generated a heat map showing pairwise sequence identity

224   of 1 kb along the centromeres (Supplementary Fig. 11a), and observed that the *CentO*

225   sequences had the highest similarity in the middle and declined to both sides

226   (Supplementary Fig. 11a). Furthermore, the profile of *CentO* sequences

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

228 genomic level.

229

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

231 Although several linear rice pan-genome had been constructed, the sequence was

232 mainly based on *de novo* assembly of short-read re-sequencing data (Wang et al.,

233 2018). In addition to 66 short-reads assembled genomes (Zhao et al., 2018), 13

234 genomes assembled by long-reads were selected to construct pan-genome

235 (Supplementary Table 29). The above 79 rice varieties (7 *O. sativa aus*, 27

236 *indica/xian*, 25 *temperate japonica/geng*, 6 *tropical japonica/geng*, 1 *O. sativa*

237 *aromatic* and 13 *O. rufipogon*) represent the major of *O. sativa* and *O. rufipogon*

238 groups (Supplementary Table 29). Phylogenetic tree was constructed by using jacard

239 similarity between long-kmer datasets to determine the similarity between different

240 genomes. From the phylogenetic relationship, it was obvious that the same subgroups

241 of Asian cultivated rice were clustered together, including *temperate japonica/geng,*

242 *tropical japonica/geng*, *indica/xian* and *aus* (Fig 3a). It can also be observed that *xian*

243 and *geng* were close to different subgroups of wild rice (Wing et al., 2018a; Xie et al.,

244 2020). ZS97 and MH63 were in different branches in the *O. sativa xian* subgroup (Fig

245 3a). Previous studies had divided them into the *indica/xian II* and *indica/xian I*

246 subgroups respectively, which represented different *O. sativa indica* population and

247 showed a large genetic difference (Xie et al., 2015). We used the gap-free genome

248 MH63RS3 as the reference to identify presence/absence variations (PAVs) in other

249 rice varieties to construct graph-based pan-genome of *O. sativa* (Os-GPG), which can

250 not only identify complex SVs, but also improve the accuracy of variation calls

251 around SVs (Rakocevic et al., 2019; Liu et al., 2020). After filtering redundancy and

252 decontamination, the pan-PAVs of Asian cultivated rice is ~320 Mb, of which *xian* is

253 169 Mb and *geng* is 145 Mb (Fig. 3a; Supplementary Table 30). Affected by the

254 diversity of *xian* rice, the PAV of *xian* rice was greater, even when the *xian* genome

255 was used as the reference. 17,365 protein-coding genes were annotated in pan-PAVs

256   of Asian cultivated rice that were not present in reference genome (Supplementary

257   Table 30). We merged 454,187 PAVs from all genomes into a set of 278,567

258   nonredundant PAVs. Further, vg toolkit (Garrison et al., 2018) was used to construct a

259   graph-based pan-genome of rice, which can be directly used for read mapping and

260   GWAS analysis (Fig. 3b). It is the first graph-based pan-genome obtained from a

261   gap-free reference genome in rice. The pan-PAVs sequence had a lower gene density

262   than reference, but contained abundant resistance genes (NLRs). We identified 557

263   NLRs in pan-PAVs, and this number is similar to the reference genomes (MH63:509;

264   Nip: 473 (Wang et al., 2019)) (Supplementary Table 31). Therefore, when a single

265   reference genome was used to study the adaptability of rice, almost half of the NLR

266   genes are missed. A large number of NLRs were imbalanced in *'xian'* and *'geng'*

267   subgroups, and some NLRs only existed in a few wild rice varieties (Supplementary

268   Fig. 12b). The Os-PGP provides valuable resources and should promote rice studies in

269   the post-genomic era. The distribution of PAVs and NLRs of ZS97 and MH63 were

270   similar in other chromosomes, while highly different in the end of chromosome 11

271   (Fig. 3c, Supplementary Fig. 12a). In this region, we found two large SVs, named

272   MH-Ex1 and MH-INS1, between ZS97 and MH63 (Supplementary Fig. 13a).

273   Through mapping the PacBio HiFi reads of ZS97 and MH63 to the end of

274   chromosome 11 of MH63RS3 genome, we clearly observed the two large SVs. The

275   reads of MH63 can continuously span these two regions, while ZS97 reads cannot

276   cover these regions (Fig. S11b). For MH-Ex1, most of the resistance genes in ZS97

277   amplified 2-10 times in MH63 (Fig. 3d; Supplementary Table 32), resulting a large

278   genomic sequence expansion (from 0.18 Mb in ZS97 to 0.82 Mb in MH63). It is very

279   interesting that most of the expanded resistance genes are not expressed or lowly

280   expressed in most tissues except root (Fig. 3d; Supplementary Fig. 13c;

281   Supplementary Table 32). For MH-INS1, MH63RS3 genome had an 857 kb insertion

282   compared with ZS97RS3 genome, including eleven resistance genes with low

283   expression levels in most tissues except root (Supplementary Table 33). We further

284   scanned the two SVs (MH-Ex1 and MH-INS1) in the remaining 25 rice genomes

285 assembled based on PacBio long-read sequencing, and observed that MH-Ex1 and

286 MH-INS1 were incomplete in all the other rice varieties compared with MH63

287 genome (Zhou et al. 2020) (Fig. 3d, Supplementary Fig. 14; Supplementary Table 34).

288 The above example indicated the genetic advantage of MH63 a donor of resistance

289 genes. This was an illustration that Os-PGP will provide the full range of short to

290 long-range SVs that exist across the *O. sativa*.

291 In summary, we assembled two gap-free genomes of *xian* rice ZS97 and MH63,

292 which are the first report of gapless plant genomes up to now. Based on these

293 genomes, we analyzed and compared the complete centromeres of all chromosomes in

294 both rice varieties, and observed that >1,500 genes were existed in centromere regions

295 and ~16% of them were actively expressed. Based on the gap-free MH63RS3 genome,

296 a graph-based rice pan-reference-genome was constructed containing

297 presence/absence variations of 79 rice varieties, which can be used as a solid

298 foundation for further genome wide association studies.

299 **METHODS**

300 **Plant Materials and Sequencing**

301 Fresh young leaf tissue was collected from *O. sativa* ZS97 and MH63 plants. We

302 constructed SMRTbell libraries as described in previous study (Pendleton, M. et al.

303 2015). The genomes of MH63 and ZS97 were sequenced using PacBio Sequel II

304 platform (Pacific Biosciences), including 8.34 Gb HiFi reads (~23x coverage) and

305 48.39Gb CLR reads (~131x coverage) for ZS97, and 37.88 Gb HiFi reads (~103x

306 coverage) and 48.97 Gb CLR reads (~132x coverage) for MH63 genomes. Plant

307 tissues were extracted using the BioNano plant tissue extraction protocol. We

308 embedded the extracted DNA in BioRad LE agarose for subsequent washes of TE,

309 proteinase K (0.8mg/ml), and RNAse A (20μL/mL) treatments in lysis buffer. The

310 Agarose plugs were then melted using agarase (0.1 U/μL, New England Biolabs) and

311 dialyzed on millipore membranes (0.1μm) with TE to equilibrate ion concentrations.

312 We then nicked the DNA with a nickase restriction enzyme BssSI (2U/μL) with a 6 bp

313 sequence recognition motif. Labeled nucleotides were incorporated at breakpoints and

314 the DNA was counterstained. Each sample was loaded onto 2 nanochannel flow cells

315 of an Irys machine for DNA imaging. Truseq Nano DNA HT Sample preparation Kit

316 following manufacturer's standard protocol (Illumina) was used to generate the

317 libraries for Illumina paired-end genome sequencing. These libraries were sequenced

318 to generate 150 bp paired-end reads by Illumina HiSeq X Ten platform with 350 bp

319 insert size.

320

321 **Genome Assembly and Assessment**

322 In this work, seven tools based on different algorithms were performed to assemble

323 the genomes of ZS97 and MH63. (1) Canu v1.8 (Koren S et al., 2017) was used to

324 assemble the genomes with default parameters; (2) FALCON toolkit v0.30 (Carvalho

325 et al., 2016) was applied for assembly with the parameters: pa_DBsplit_option =

326 -s200 -x500, ovlp_DBsplit_option = -s200 -x500, pa_REPmask_code =

327 0,300;0,300;0,300, genome_size = 400000000, seed_coverage = 30, length_cutoff =

328 -1, pa_HPCdaligner_option =-v -B128 -M24, pa_daligner_option=-k18 -w8 -h480

329 -e.80 -l5000 -s100, falcon_sense_option=--output-multi --min-idt 0.70 --min- cov 3

330 --max-n-read 400, falcon_sense_greedy=False, ovlp_HPCdaligner_option=-v -M24

331 -l500, ovlp_daligner_option=-h60 -e0.96 -s1000, overlap_filtering_setting=--max-diff

332 100 --max-cov 100- -min-cov 2, length_cutoff_pr=1000; (3) MECAT2 (Xiao et al.,

333 2017) was utilized to assemble with the parameters: "GENOME_SIZE=400000000,

334 MIN_READ_LENGTH=2000, CNS_OVLP_OPTIONS="", CNS_OPTIONS="-r 0.6

335 -a 1000 -c 4 -l 2000", CNS_OUTPUT_COVERAGE =30,

336 TRIM_OVLP_OPTIONS="-B", ASM_OVLP_OPTIONS="-n 100 -z 10 -b 2000 -e

337 0.5 -j 1 -u 0 -a 400", FSA_OL_FILTER_OPTIONS = "--max_overhang = -1

338 --min_identity = - 1", FSA_ASSEMBLE_OPTIONS = "", GRID_NODE = 0,

339 CLEANUP = 0, USE_GRID = false "; (4) Flye 2.6-release (Kolmogorov et al., 2019)

340 was set with "--genome-size 400m"; (5) Wtdbg2 2.5 (Ruan et al., 2020) was used to

341 assemble with parameters " -x sq, -g 400m", and then Minimap2 (Li 2018) was

12

342 employed to map the PacBio CLR data to the assembly results, and wtpoa was

343 utilized polish and correct the wtdbg2 assembly results; (6) NextDenovo v2.1-beta.0

344 (https://github.com/Nextomics/NextDenovo) was applied for assembly with

345 parameters "task = all, rewrite = yes, deltmp = yes, rerun = 3, input_type = raw,

346 read_cutoff = 1k, seed_cutoff = 44382, blocksize = 2g, pa_correction = 20,

347 seed_cutfiles = 20, sort_options = -m 20g -t 10 -k 40, minimap2_options_raw = -x

348 ava-ont -t 8, correction_options = -p 10, random_round = 20, minimap2_options_cns

349 = -x ava-pb -t 8 -k17- w17, nextgraph_options = -a 1"; (7) Miniasm-0.3-r179 (Heng

350 Li 2016) with default parameters. Based on these seven software, Genome Puzzle

351 Master GPM (Zhang et al., 2016) was performed to integrate and optimize the

352 assembled contigs, and visualize the complete chromosomes. Based on the HiFi and

353 CLR sequencing data, we used GenomicConsensus package of

354 SMRTLink/7.0.1.66975 (https://www.pacb.com/support/) to polish the assembled

355 genome twice with Arrow algorithm, the parameters are: --algorithm=arrow. Pilon

356 (Walker et al., 2014) was used for polishing the genomes based on Illumina data with

357 the parameters: --fix snps, indels. This process repeated twice. Molecules were then

358 assembled using Bionano IrysSolve pipeline

359 (https://bionanogenomics.com/support-page/) to create optical maps. Images were

360 interpreted quantitatively using Bionano AutoDetect 2.1.4.9159 and data was

361 visualized using IrysView v2.5.1. These assemblies were used with draft genome

362 assemblies to validate and scaffold the sequences. Bionano map data was aligned to

363 the merged contigs using RefAlignerAssembler in IrysView software to do the

364 verification.

365 ZS97RS3 and MH63RS3 genome completeness assessment using BUSCO v4.0.6

366 (Felipe A et al., 2015). Besides, we mapped the PacBio HiFi reads and PacBio CLR

367 reads (using Minimap2 (Li 2018)), Illimina reads (using BWA-0.7.17 (Jo H et al.,

368 2015)), BES/BAC reads (using BLASTN v2.7.1 (Altschul et al., 1990)), Hi-C reads

369 (using HiC-Pro v2.11.1 (Servant et al., 2015)), RNA-Seq reads (using Hisat2 v2.1.0

370 (Kim et al., 2015)) to both genome assemblies find both assemblies performed well.

371

Gene and Repeat Annotations

373 MAKER-P (Campbell et al., 2014) version 3 was used to annotate the ZS97RS3 and

374 MH63RS3 genomes. All the evidences are the same as that used for RS1 genomes. To

375 ensure the consistency of RS1 version, genes can completely map to RS3 genome

376 were retained. The new genes in gap regions were obtained from MAKER-P

377 (Campbell et al., 2014). Genes encoding transposable elements were identified and

378 transitively annotated by searching against the MIPS-REdat Poaceae version 9.3p

379 (Nussbaumer et al., 2013) database using TBLASTN (Altschul et al., 1990) with

380 E-value 1e-10. tRNAs were identified with tRNAscan-SE (Lowe, T. M. & Eddy, S. R.

381 19997) using default parameters; rRNA genes were identified by searching the

382 genome assembly against the rRNA sequences of Nipponbare using BLASTN v2.7.1

383 (Altschul et al., 1990); miRNAs and snRNAs were predicted using INFERNAL of

384 Rfam (Griffiths-Jones, S. et al., 2005) (v14.1). Repeats in the genome were annotated

385 using RepeatMasker (Smit et al. 2015) with RepBase (Bao et al., 2015), TIGR Oryza

386 Repeats (v3.3) with RMBlast search engine. For the overlapping repeats in different

387 classes, LTR retrotransposons were kept first, next TIR, and then SINE and LINE,

388 finally helitrons. This priority order was based on stronger structural signatures.

389 Besides, the known nested insertions models (LTR into helitron, helitron into LTR,

390 TIR into LTR, LTR into TIR) were retained. The identified repetitive elements were

391 further characterized and classified using PGSB repeat classification schema.

392 LTR_FINDER (Xu Z, Wang H 2007) was used to identify complete LTR-RTs with

393 target site duplications (TSDs), primer binding sites (PBS) and polypurine tract

394 (PPT).

395

Chromatin Immunoprecipitation (ChIP) and ChIP-seq

397 The procedures for chromatin immunoprecipitation (ChIP) were adopted from Nagaki

398 et al. (2003) and Walkowiak et al. (2020). The nuclei were isolated from 4-week-old

399 seedlings. The nuclei were digested with micrococcal nuclease (Sigma-Aldrich, St.

400 Louis, MO) to liberate nucleosomes. The digested mixture was incubated overnight

401 with 3 µg of rice CENH3 antibody at 4°C. The target antibodies were captured from the

402 mixture using Dynabeads Protein G (Invitrogen, Carlsbad, CA). ChIP-seq libraries

403 were constructed using TruSeq ChIP Library Preparation Kit (Illumina, San Diego, CA)

404 following the manufacturer's instructions and the libraries were sequenced on Illumina

405 HiSeqX10 with 2x150 bp sequencing run.

406

Fluorescence *in situ* Hybridization (FISH)

407

408 *Slide Preparation*

409 Mitotic chromosomes were prepared as described by Koo and Jiang (2009) with

410 minor modifications. Root tips were collected from plants and treated in a nitrous oxide

411 gas chamber for 1.5 h. The root tips were fixed overnight in ethanol:glacial acetic acid

412 (3:1) and then squashed in a drop of 45% acetic acid.

413 *Probe Labeling and Detection*

414 The ChIPed DNAs were labeled with digoxigenin-16-dUTP using a nick translation

415 reaction. The clone, maize 45S rDNA (Koo and Jiang 2009) was labeled with

416 biotin-11-dUTP (Roche, Indianapolis, IN). Biotin- and digoxigenin-labeled probes

417 were detected with Alexa Fluor 488 streptavidin antibody (Invitrogen) and

418 rhodamine-conjugated anti-digoxigenin antibody (Roche), respectively.

419 Chromosomes were counterstained with 4',6-diamidino-2-phenylindole (DAPI) in

420 Vectashield antifade solution (Vector Laboratories, Burlingame, CA). The images

421 were captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy LLC,

422 Thornwood, NY) using a cooled CCD camera CoolSNAP HQ2 (Photometrics,

423 Tucson, AZ) and AxioVision 4.8 software. The final contrast of the images was

424 processed using Adobe Photoshop CS5 software.

425

The Completeness of Centromeres on MH63RS3 and ZS97RS3 Chromosomes

426

427 Based on the final RS3 genomes, we use BLAST (Altschul et al., 1990) to align the c

428 *CentO* satellite repeats in rice to the reference genome with E-value 1e-5, then use

429 BEDtools (Quinlan et al., 2014) to merge the result with the parameter -d 50000. Then,

430 from the outside to the inside, if the number of consecutive *CentO* is less than 5, it is

431 classified as core region if the number of consecutive *CentO* is greater than 5 but less

432 than 10, and the distance between two *CentO* clusters a less than 10kb, it is classified

433 as core region; if the number of consecutive *CentO* is more than 10, it is directly

434 classified as core region.

435   For the identification of the whole centromere region, we use BWA-0.7.17 (Jo H

436 et al., 2015) to align the CENH3 ChIP-Seq reads to MH63RS3 and ZS97RS3

437 genomes, and use SAMtools (Li H et al., 2009) to filter the results with mapQ value

438 above 30; then we use MACS2 (Zhang Y et al., 2008) to call the peaks of CENH3.

439 Finally, we combined the distribution of CENH3 histone, CentOS, repeats and genes

440 to jointly define all the centromeric region of MH63RS3 and ZS97RS3 genomes. It

441 should be noted that when determining the peaks of CENH3 histones, the standard is

442 that if three consecutive peaks value > 30 and no cluster interference, the last peak

443 position is defined as the centromeric boundary position; if three consecutive peaks

444 value > 30 but there has cluster interference, reduce the peak value standard to 20, and

445 then define the centromere boundary; combined with manual adjustment of the

446 position.

447   To compare of *CentO* sequence similarity, first we use BEDtools (Quinlan et al.,

448 2014) to obtain sequences of centromere core regions, and divide them into 1 kb

449 continuous sequences; then we use Minimap2 (Li 2018) to align the sequences, the

450 parameters are: -f 0.00001 -t 8 -X --eqx -ax ava -pb; finally, we use a custom python

451 script to filter the result file, and use R to generate a heat map showing pairwise

452 sequence identity (Logsdon, Vollger et al. 2020).

453

454 **Telomere Sequence Identification**

455 The telomere sequence 5'-CCCTAAA-3' and the reverse complement of the seven

456 bases were searched directly. In addition, we used BLAT (Kent WJ 2002) to search

457 telomere-associated tandem repeats sequence (TAS) from TIGR Oryza Repeat

458 database (Ouyang et al., 2004) in whole genome.

459

460 **Identification of PAVs**

461 We selected 79 rice varieties to construct phylogenetic tree, 66 were from previous

462 studies (Zhao et al., 2018) and 11 were downloaded from NCBI (as of 1-30-2020).

463 Sourmash was used to compute hash sketches from genome sequences (k-mer = 301)

464 and calculate jaccard similarity of 79 rice genomes to generate phylogenetic tree

465 (Pierce et al., 2019). The rice genomes were aligned to reference genome MH63 using

466 Mummer(4.0.0beta2) (Marçais et al., 2018) with parameters settings '-c 90 -l 40'.

467 Then used "show-diff" to select for unaligned regions. Further we merged all *O.*

468 *sativa indica* and *O. sativa japonica* unaligned sequences and then used

469 CD-HIT(v4.8.1) (Fu et al., 2012) to remove redundant sequences. Finally, we used

470 blastn to remove contaminate sequences with parameters settings '-evalue 1e-5

471 -best_hit_overhang 0.25 -perc_identity 0.5 -max_target_seqs 10' and the rest is PAVs

472 sequences.

473

474 **Prediction of NLR Genes**

475 We first predicted domains of genes with InterProScan (Jones et al., 2014), which can

476 analyze peptide sequences against InterPro member databases, including ProDom,

477 PROSITE, PRINTS, Pfam, PANTHER, SMART and Coils. Pfam and Coils were used

478 to prediction NLRs. NLRs were defined to contain at least NB, a TIR, or a

479 CCR(RPW8) domain and we classified NLRs based on above structural features.

480 NLRs domain contain only NB (Pfam accession PF00931), TIR (PF01582), RPW8

481 (PF05659), LRR (PF00560, PF07725, PF13306, PF13855) domains, or CC motifs

482 (Van de Weyer et al., 2019).

483

484 **Identification of Collinear Orthologues**

485 MCscan (python version) (Tang et al., 2008) was used to identify collinear

17

486   orthologues between chromosome 11 of ZS97RS3 and MH63RS3 genomes with

487   default parameters.

488

Construction of Graph-based Pan-genome

490   MH63RS3 was set as a reference and the pan-PAVs sequences were saved in variant

491   call format (VCF). The graph-based pan-genome was construct via the vg

492   (https://github.com/vgteam/vg, version v1.29.0) toolkit (Garrison et al., 2018) with

493   default parameters.

494

DATA AVAILABILITY

496   All the raw sequencing data generated for this project are achieved at NCBI under

497   accession numbers SRR13280200, SRR13280199 and SRR13288213 for ZS97,

498   SRX6957825, SRX6908794, SRX6716809 and SRR13285939 for MH63. The

499   genome assemblies are available at NCBI (CP056052-CP056064 for ZS97RS3,

500   CP054676-CP054688 for MH63RS3) and annotations are visualized with Gbrowse at

501   http://rice.hzau.edu.cn. All the materials in this study including introgression lines are

502   available upon request.

AUTHOR CONTRIBUTIONS

510   L.-L.C., J.Z., R.W. and Q.Z. designed studies and contributed to the original concept

511   of the project. J.P. and D.-H.K. performed the ChIP-seq and FISH experiments. D.K.,

512   E.L., S.L., J.T., D.Y., J.U. and R.W. performed the genome and BioNano sequencing.

513   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

514   genome assembling and annotation, comparative genomics analysis and other data

515   analysis. J.-M.S., W.-Z.X., S.W., J.P., D.-H.K., L.-L.C. and J.Z. wrote the paper.

516   W.X., R.W. and Q.Z. contributed to revisions.

517

523

524   **ONLINE CONTENT**

525   Any methods, additional references, Research reporting summaries, source data,

526   statements of code and data availability and associated accession codes are available

527   online.

528   **REFERENCES**

529   **Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.** (1990). Basic
530   local alignment search tool. J. Mol. Biol. **215:**403–410.

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

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

535   **Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R.,**
536   **and Jiang, J.** (2002). Functional rice centromeres are marked by a satellite repeat
537   and a centromere-specific retrotransposon. Plant Cell **14:**1691–1704.

538   **Simão, F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., and Zdobnov,**
539   **E.M.** (2015). BUSCO: assessing genome assembly and annotation completeness
540   with single-copy orthologs. Bioinformatics **31:**3210–3212.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

584 **Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z.,**
585 **Shi, M., et al.** (2020). Pan-genome of wild and cultivated soybeans. Cell
586 **182:**162–176.

587 **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

672 between ZS97RS3 and MH63RS3 were linked as the gray lines. All the RS1 gap

673 regions were closed in RS3 and showed in the yellow block. The black triangle

674 indicated the telomere, there are 7 T-to-T chromosomes in ZS97RS3 (Chr01, Chr02,

675 Chr03, Chr04, Chr06, Chr07, Chr11) and 10 T-to-T chromosomes in MH63RS3

676 (Chr01, Chr02, Chr03, Chr04, Chr05, Chr06, Chr07, Chr09, Chr10, Chr12). All the

677 centromeres are complete and repeat length distribution diagrams were plotted

678 above/under each chromosome; b). Histogram showed the reads coverage for different

679 libraries in MH63RS3 and ZS97RS3, including BAC, CCS and CLR reads.

680 **Fig. 2│Complete rice centromeres.**

681 **a**, The definition of MH63RS3 centromere. the first to fourth layers indicate the

682 histone CENH3 Chip-seq distribution, the *CentO* satellite distribution, t genes

683 distribution, and of TE distribution, respectively. The dotted frame represents the final

684 centromere area. **b**, FISH signals detected in metaphase of meiosis for MH63RS3 and

685 ZS97RS3, white arrows indicate DNA elements in the centromeric region. **c**,

686 Coverage of HiFi, CLR, Illumina reads and distribution of TEs in the centromere on

687 Chr01 (extended 500 kb left and right) of MH63RS3. **d**, Characteristics of the

688 centromere on Chr01 of MH63RS3. The first layer is histone CENH3 distribution, the

689 second layer is the CentOS distribution, the third layer is the Genes distribution, the

690 fourth to sixth levels are gene expression, the seventh to ninth levels are methylation

691 distribution, the tenth layer is CentOS sequence similarity.

692 **Fig. 3│The graph-based pan-genome and pan-NLRome of rice.**

693 **Figure 3. a,** Phylogenetic tree of the 79 rice varieties. 79 rice varieties phylogenetic

694 tree (left), black represents wild rice varieties, orange represents *O. sativa aus*, Orange

695 shadow represents *O. sativa indica*, blue shadow represents *O. sativa japonica*, heat

696 map represents the jaccard similarity of pairwise rice (middle), and bar graph

697 represents the number of PV per rice (right). **b,** The schematic diagram of rice

698 graph-based pan-genome. **c,** Distribution of the difference regions between ZS97RS3

23

699   and MH63RS3 on the chromosome. **d,** The expansion structural variation of

700   MH63RS3. The expansion structural variation at the end of chromosome 11 of

701   MH63RS3, from top to bottom are the gene collinearity of ZS97RS3 and MH63RS3,

702   the TE distribution, the gene expression in this region and coverage ratio of two

703   structural variations in 25 rice varieties.

a

b

Os-PGP

Gap-free RefSeq

Deletion

Insertion

Individual genomes

Reads mapping to RefSeq   Reads mapping to variant region   Core genome   Dispensable genome

c

Chromosome length (Mb)

MH63 presence ratio   ZS97 presence ratio

0   1   0   1

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   [0 - 50]
Root   [0 - 50]
Leaf   [0 - 50]

SVs   Accessions of Asian rice   synteny region (%)

MH-Ex1

MH-INS1

Os127742   IR8   Tetep   Azucena   IRGC106162   Os125927   ZSRS3   Kitaake   Nipponbare   OsIR64   Os127652   R498   Carolina   R600   Os132424   9311   OnivRS2   HR-12   OR_W1943   Os127564   Os127518   Os117425   Os128077   Os125619   Os132278

Jaccard similarity of sigatures

0.0   0.2   0.4   0.6   0.8   1.0

Number of PVs

0   5000   10000

169 Mb Presence

145 Mb Presence

indica

japonica

aus

indica

temperate japonica

tropical japonica

Or-I

Or-III