

1 **Assembly and Validation of Two Gap-free Reference**  
2 **Genomes for *Xian/indica* Rice Reveals Insights into Plant**  
3 **Centromere Architecture**

4

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

10

11 <sup>1</sup>National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University,  
12 Wuhan, 430070, China

13 <sup>2</sup>College of Life Science and Technology, Guangxi University, Nanning, 530004, China

14 <sup>3</sup>Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

15 <sup>4</sup>Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson,  
16 Arizona 85721, USA

17 <sup>5</sup>Center for Desert Agriculture, Biological and Environmental Sciences & Engineering  
18 Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal,  
19 23955-6900, Saudi Arabia

20 <sup>6</sup>Plant and Wildlife Science Department, Brigham Young University, Provo, UT 84602, USA

21 <sup>7</sup>International Rice Research Institute (IRRI), Strategic Innovation, Los Baños, 4031 Laguna,  
22 Philippines

23 <sup>8</sup>These authors contributed equally to this work.

24 **\*Correspondence: Jesse Poland ([jpoland@ksu.edu](mailto:jpoland@ksu.edu))**

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

26 **Ling-Ling Chen ([llchen@mail.hzau.edu.cn](mailto:llchen@mail.hzau.edu.cn))**

27 **ABSTRACT**

28 **Rice (*Oryza sativa*), a major staple throughout the world and a model system for**  
29 **plant genomics and breeding, was the first crop genome completed almost two**  
30 **decades ago. However, all sequenced genomes to date contain gaps and missing**  
31 **sequences. Here, we report, for the first time, the assembly and analyses of two**  
32 **gap-free reference genome sequences of the elite *O. sativa xian/indica* rice**  
33 **varieties ‘Zhenshan 97 (ZS97)’ and ‘Minghui 63 (MH63)’ that are being used as a**  
34 **model system to study heterosis. Gap-free reference genomes also provide global**  
35 **insights into the structure and function of centromeres. All rice centromeric**  
36 **regions share conserved centromere-specific satellite motifs but with different**  
37 **copy numbers and structures. Importantly, we demonstrate that >1,500 genes**  
38 **are located in centromere regions, of which ~15.6% are actively transcribed. The**  
39 **generation and release of both the ZS97 and MH63 gap-free genomes lays a solid**  
40 **foundation for the comprehensive study of genome structure and function in**  
41 **plants and breed climate resilient varieties for the 21<sup>st</sup> century.**

42

43 **Key words:** gap-free genome, ZS97, MH63, centromere architecture

44

45 **INTRODUCTION**

46 *Oryza sativa* ‘*xian/indica*’ and ‘*geng/japonica*’ groups, previously subsp. *indica* and  
47 subsp. *japonica* respectively, are two major types of Asian cultivated rice (Wang et al.,  
48 2018). *Xian* varieties are broadly studied as they contribute over 70% of rice  
49 production worldwide and are genetically more diverse than *geng* rice. Over the past  
50 30 years, two *xian* varieties Zhenshan 97 (ZS97) and Minghui 63 (MH63) have  
51 emerged as important model system in rice breeding and genomics being the parents  
52 of the elite hybrid Shanyou 63 (SY63), historically the most widely cultivated rice  
53 hybrid in China. Understanding the biological mechanisms behind the elite  
54 combination of ZS97 and MH63 to form the SY63 hybrid is foundational to help  
55 unravel the mystery of heterosis (Yu et al., 1997; Hua et al., 2002; Hua et al., 2003;  
56 Huang et al., 2006; Zhou et al., 2012); Further, ZS97 and MH63 represent two major  
57 varietal subgroups in *xian* rice as they show many complementary agronomic traits,  
58 and a number of important genes have been cloned based on genetic populations  
59 generated using these two varieties as parents (Sun et al., 2004; Fan et al., 2006; Xue  
60 et al., 2008). Although we previously generated two reference genome assemblies  
61 ZS97RS1 and MH63RS1 in 2016, approximately 10% of each genome remained  
62 unassembled/unplaced (Zhang et al., 2016a). Upon further analysis and editing we  
63 were able to fill the majority of gaps in each assembly and released upgraded versions  
64 of these two assemblies in 2018 (<http://rice.hzau.edu.cn>), yet eight (ZS97) and seven  
65 (MH63) gaps still remained.

66 To bridge all remaining assembly gaps across each genome we incorporated  
67 high-coverage and accurate long-reads sequence data and multiple assembly strategies  
68 to successfully generate two gap-free genome assemblies of *xian* rice ZS97 and  
69 MH63, the first gap-free plant genome assemblies publicly available to date.  
70 Importantly, we had the first opportunity to study and compare the centromeres of all  
71 chromosomes side by side across both rice varieties. More than expected, >1,500

72 genes were identified in rice centromere regions, ~15.6% of which were found to be  
73 actively transcribed.

74

## 75 **RESULTS**

### 76 **Assembly and Validation of Gap-free Reference Genome Sequences for ZS97 and** 77 **MH63**

78 In this project, 56.73 Gb (~150X) and 86.85 Gb (~230X coverage) of PacBio reads  
79 (including both HiFi and CLR modes) were generated for ZS97 and MH63,  
80 respectively, using the PacBio Sequel II platform ([Supplemental Figure 1](#),  
81 [Supplemental Table 1](#)). The PacBio HiFi and CLR reads were assembled separately  
82 with multiple *de novo* assemblers including Canu ([Koren et al., 2017](#)), FALCON  
83 ([Carvalho et al., 2016](#)), MECAT2 ([Xiao et al., 2017](#)) *etc.* (see Methods), and then the  
84 assembled contigs were merged with the two upgraded assemblies using Genome  
85 Puzzle Master (GPM) ([Zhang et al., 2016b](#)) ([Supplemental Table 2-3](#)). Finally, two  
86 gap-free reference genomes were produced, named as ZS97RS3 and MH63RS3,  
87 which contained 12 pseudomolecules with total lengths of 391.56 Mb and 395.77 Mb,  
88 respectively ([Figure 1a](#), [Table 1](#)). Compared with the previous bacterial artificial  
89 chromosome (BAC) based RS1 genome assemblies, the new RS3 assemblies gained  
90 ~36 to 45 Mb of additional sequence by filling 223 (ZS97RS1) and 167 (MH63RS1)  
91 gaps across both genomes ([Supplemental Table 4](#)). In addition, the new assemblies  
92 corrected a few mis-orientated or mis-assembled regions caused by reliance on the  
93 Os-Nipponbare-Reference-IRGSP-1.0 sequence as a guide to produce the RS1  
94 pseudomolecules (e.g. the 6 Mb inversion on Chr06) ([Supplemental Figure 2a-c](#),  
95 [Supplemental Table 4](#)). These anomalies could be corrected by newly assembled  
96 contigs that were long enough to span these ambiguous regions.

97 Using the 7-base telomeric repeat (CCCTAAA at 5' end or TTTAGGG at 3' end)  
98 as a probe, we identified 19 and 22 telomeres that resulted in 7 and 10

99 telomere-to-telomere (T-to-T) pseudomolecules in ZS97RS3 and MH63RS3  
100 assemblies, respectively (Figure 1a, Supplemental Table 5-6).

101 The accuracy and completeness of the RS3 assemblies were validated in multiple  
102 ways. Chromosome conformation capture sequencing (Hi-C) and Bionano optical  
103 maps showed high consistency across all pseudomolecules demonstrating correct  
104 ordering and orientation (Supplemental Figure 3, Supplemental Table 2). Genome  
105 completeness was demonstrated by high mapping rates with various raw sequences,  
106 such as raw HiFi/CLR/Illumina reads, paired BAC-end sequences, and paired-end  
107 short reads from 48 RNA-seq libraries, all of which mapped at over 99% across each  
108 assembly (Supplemental Table 7-9). The evenly distributed breakpoints of aligned  
109 short and long reads indicated that all sequence connections are of high accuracy at  
110 single-base level in these final assemblies (Supplemental Figure 4). For gene content  
111 assessment, both ZS97RS3 and MH63RS3 assemblies captured 99.88% of a BUSCO  
112 1,614 reference gene set (Supplemental Table 10). Long terminal repeat (LTR)  
113 annotation further revealed the LTR assembly index (LAI) for the ZS97RS3 and  
114 MH63RS3 assemblies were 24.01 and 22.74, respectively, which meets the standard  
115 of gold/platinum reference genomes (Ou et al., 2018, Mussurova et al., 2020) (Table  
116 1). More than 1,500 rRNAs were identified in ZS97RS3 and MH63RS3 assemblies  
117 (Supplemental Figure 5), whereas only tens were identified in the original RS1  
118 assemblies.

119

## 120 **Annotation and Comparison of Gap-free Reference Genome Sequences for ZS97** 121 **and MH63**

122 To annotate the ZS97 and MH63 RS3 assemblies for transposable element (TE) and  
123 other repetitive sequence content, we used RepeatMasker (Zhi et al., 2006) with the  
124 latest RepBase (Bao et al., 2015) and TIGR Oryza Repeats (v3.3) (Ouyang and Buell,  
125 2004) as libraries. As a result, we identified 465,242 TE sequences in ZS97RS3  
126 (181.00 Mb in total length) and 468,675 TE sequences in MH63RS3 (~182.26Mb)

127 (Supplemental Table 11-12), which accounted ~46.16% and ~45.99% of each  
128 assembly and was approximate 5% greater than that in the previous RS1 assemblies  
129 (i.e. ZS97RS1=41.28%; MH63RS3=41.58%). The repeat content increases were  
130 primarily due to the fact that over 80% of the gaps closed were in TE-rich regions  
131 (82.86% of the 45 Mb closed-gaps were TEs in ZS97RS3, and 84.17% of the 36 Mb  
132 closed-gaps were TEs in MH63RS3), and the above updated TE library.

133 Next we employed MAKER-P (Campbell et al., 2014) to annotate the ZS97RS3  
134 and MH63RS3 assemblies using the identical evidence including EST, RNA-Seq, and  
135 protein used to annotate the RS1 assemblies (Supplemental Fig. 1). In order to retain  
136 consistency across different assembly versions, 51,027 and 50,341 previously  
137 annotated gene models in the ZS97RS1 and MH63RS1 assemblies, respectively, were  
138 lifted onto the RS3 annotations. Combining models annotated with MAKER-P in the  
139 newly assembled regions, the final annotations in ZS97RS3 and MH63RS3 contained  
140 60,935 and 59,903 gene models, of which 39,258 and 39,406 were classified as  
141 non-TE gene loci (Table 1), thereby resulting in 4,648 (ZS97) and 2,082 (MH63)  
142 additional non-TE genes than previously identified in the RS1 assemblies,  
143 respectively. More than 92% of all annotated gene models were supported by  
144 homologies with known proteins or functional domains in *Oryza* and other species  
145 (Supplemental Table 13-14).

146 Based on our new assemblies, the annotation and comparative analyses of  
147 non-coding RNAs (transfer RNAs, ribosomal RNAs, small nucleolar RNAs,  
148 microRNAs) (Supplemental Figure 4), single nucleotide polymorphisms (SNPs) and  
149 insertions/deletions (InDels) among ZS97, MH63 and Nipponbare (Supplemental  
150 Figure 6, Supplemental Table 15), presence/absence variations (PAVs) (Supplemental  
151 Table 16), and genes in different categories ('identical', 'same length', 'collinear',  
152 'divergent' and 'variety-specific' genes) (Supplemental Table 17) that were previously  
153 identified in the RS1 versions were updated.

154       After comparing the PAV distribution across each chromosome of both gap-free  
155 assemblies, we noticed an abundance of structural variations (SVs) near the ends of  
156 the long-arms of chromosome 11 ([Figure 2a](#)). Two large SVs, one expansion region  
157 (30.75 – 31.57 Mb) and one insertion region (31.90 – 32.76 Mb), were uniquely  
158 detected in MH63 (hereafter named as MH-E and MH-I, respectively). Raw  
159 sequencing read alignments to these two regions clearly showed that MH-E and MH-I  
160 regions could be continuously covered by MH63 reads but only partially covered by  
161 ZS97 reads ([Supplemental Figure 7](#)). Meanwhile, previous studies showed that  
162 nucleotide-binding site leucine-rich repeat (NLR) proteins were enriched in  
163 chromosome 11 ([Rice Chromosomes 11 and 12 Sequencing Consortia, 2005](#)). Hence,  
164 we performed a genome-wide homology search for NLR or NLR-like genes in both  
165 ZS97 and MH63 RS3 assemblies ([Figure 2b](#)). When putting the PAV and NLR(-like)  
166 distribution together, we could obviously determine that both MH-E and MH-I  
167 regions have more NLR(-like) content than the corresponding region in ZS97RS3  
168 assembly (30.51 – 30.69 Mb and 30.88 – 30.94 Mb, respectively) ([Supplemental](#)  
169 [Figure 7a](#)). In the MH-E region, most of the NLR(-like) genes in ZS97 amplified 2-10  
170 times in MH63 ([Figure 2c, Supplemental Table 18](#)), and interestingly, these genes are  
171 more likely to be expressed in root than in other tissues ([Figure 2c, Supplemental](#)  
172 [Figure 7c, Supplemental Table 18](#)). In the 857-kb MH-I region, eleven NLR(-like)  
173 genes also had higher expression levels in roots than in other tissues ([Figure 2d,](#)  
174 [Supplemental Table 19](#)). We further scanned the MH-E and MH-I homologous  
175 regions in 15 additional high-quality reference genomes ([Zhou et al., 2020](#)), and  
176 unexpectedly, none of them had both complete MH-E and MH-I at the same time  
177 ([Figure 2e, Supplemental Figure 8, Supplemental Table 20](#)). This unique genomic  
178 characteristic of MH63 could partially, at least, potentially explain its high resistance  
179 to blast disease.

180

## 181 **Location and Analyses of Rice Centromeres**

182 Centromeres are essential for maintaining the integrity of chromosomes during cell  
183 division, and ensure the fidelity of their inheritance. Unfortunately, until now,  
184 centromeres have remained largely under-explored, especially in larger genomes  
185 (Perumal et al., 2020). To functionally identify the location and sequence of  
186 centromeres in our gap-free genomes, we used the rice CENH3 antibody to  
187 immunoprecipitate chromatin from rice nuclei and then sequenced Illumina  
188 sequenced the captured DNA fragments (i.e. ChIP-Seq) (Figure 3a-b). To visually  
189 confirm the specificity of our ChIP experiments, we used fluorescent *in situ*  
190 hybridization (FISH) of ChIPed DNA on MH63 and ZS97 metaphase chromosomes,  
191 the results of which showed strong signals at the centromere for each chromosome  
192 (Figure 3b).

193 Using MH63RS3 as the reference, for the first time, we delimited the boundaries  
194 of each centromere and determined that the size of rice centromeres varied from 0.8  
195 Mb to 1.8 Mb (Supplemental Figure 9, Supplemental Table 21-22). We then classified  
196 rice centromeres into core and shell regions. Core centromere regions (CCRs) were  
197 identified by sequence homology to the 155-165 bp centromere-specific (*CentO*)  
198 satellite repeats which all showed high levels of CENH3 binding (Cheng et al., 2002),  
199 while shell regions were determined by the ChIP-seq signals. The lengths of the CCRs  
200 ranged from 76 kb to 726 kb in different chromosomes with a total length of 3.47 Mb  
201 in MH63RS3 (Supplemental Figure 9, Supplemental Table 21). We manually checked  
202 the entire length of each centromere region (especially the boundary regions) of both  
203 MH63RS3 and ZS97RS3 and found that the HiFi/CLR reads were evenly mapped  
204 with no ambiguous breakpoints (Figure 3c, Supplemental Figure 10), which provides  
205 strong evidence that each of the 12 centromeres in both gap-free reference genomes  
206 were contiguous and of high quality.

207 Analysis results across all centromeres in both assemblies showed that CCRs  
208 contained <130 genes in each genome but a large amount of *CentO* satellite sequences  
209 (Supplemental Figure 11), while the shell regions contained >1,400 genes, of which



210 ~16% had evidence of transcription, which included many centromere-specific  
211 retrotransposon sequences ([Supplemental Tables 23-25](#)). For example, the centromere  
212 of MH63RS3 Chr01 is 1.6 Mb, which contained a 726-kb CCR composed of 3,228  
213 *CentO* sequences and 48 genes, while the shell regions, flanking both sides of the  
214 CCR, contained 114 *CentO* sequences and 146 genes ([Figure 3d, Supplemental Table](#)  
215 [18, Supplemental Table 23](#)). For the genes located in CCRs of 12 chromosomes (109  
216 in ZS97, 129 in MH63), ~10% (7 in ZS97, 13 in MH63) were found to be transcribed  
217 in the tissues and conditions tested ([Supplemental Tables 24-25](#)). Further, of the 1,446  
218 (ZS97) and 1,764 (MH63) genes annotated in the shell regions, ~16% were found to  
219 be transcribed (231 in ZS97 and 282 in MH63). In total, 72% of gene families were  
220 shared in centromere regions of ZS97 and MH63 ([Supplemental Figure 9d](#)). Genes in  
221 the centromeres on the same chromosomes of ZS97 and MH63 were relatively  
222 conserved (mainly distributed in shell regions), an example of gene collinearity  
223 between chromosome 1 centromeres between MH63 and ZS97 was shown in [Figure](#)  
224 [3e](#). This conservation could be extended throughout the population structure (K=15)  
225 of cultivated Asian rice where the average ratio of conserved genes was ~87%,  
226 especially across the Chr05, Chr09 and Chr12 centromeres ([Supplemental Table 26](#)).

227 Gene ontology (GO) analysis showed that genes with the GO terms ‘transcription  
228 from RNA polymerase III promoter’, ‘nucleic acid binding’ and ‘nucleoplasm part’,  
229 were significantly enriched in ZS97 and MH63 centromere regions ([Supplemental](#)  
230 [Figure 10b-c, Supplemental Table 27-28](#)). Overall, these GO terms tend to have  
231 similar functions ([Supplemental Figure 12](#)). However, GO terms among different  
232 chromosomes of the same variety showed great difference, e.g., the average  
233 overlapping ratio was 37% in MH63 ([Supplemental Table 29-30](#)). We also found that  
234 the methylation levels of CG and CHG in the centromeric regions were two-fold  
235 higher than that of the whole genome ([Supplemental Table 31](#)). This phenomenon was  
236 particularly prominent in *CentO* clustered regions.

237 Based on the complete centromere location, we observed that the centromeric  
238 regions had slightly lower depth of mapped raw sequence reads than non-centromeric  
239 regions, which may be caused by highly repetitive elements; however, the lengths of  
240 those reads in centromeric and non-centromeric regions were broadly in line with  
241 each other (Supplemental Figure 11b). Detailed sequence analysis revealed an  
242 abundance of TEs in the centromeric region accounting for 78-80% of the functional  
243 centromere (Supplemental Table 32-33). In particular, the proportion of LTR/gypsy  
244 TEs, accounting for over 90% of the repeat content, is extremely higher than other  
245 types of TEs (Supplemental Figure 11c), which is an obvious barrier for fully  
246 assembling a centromere region.

247 To better understand the long-range organization and evolution of the CCRs, we  
248 generated a heat map showing pairwise sequence identity of 1-kb along the  
249 centromeres (Supplemental Figure 13a), and observed that the *CentO* sequences had  
250 the highest similarity in the middle and declined to both sides (Supplemental Figure  
251 13a). Furthermore, the profile of *CentO* sequences (Supplemental Figure 13b)  
252 illustrated the conservation of rice centromeres on the genomic level.

253 To determine if the centromere architecture found in ZS97 and MH63 was  
254 conserved among other Asian rice accessions, we compared the ZS97/MH63 CCR  
255 sequences with 15 high-quality PacBio genome assemblies that represent the  
256 population structure of cultivated Asian rice (Zhou et al. 2020). The results revealed  
257 that the lengths of *CentO* satellite repeats in the CCRs of the same chromosomes  
258 varied significantly between populations, and varieties within the same populations  
259 (Supplemental Table 34-35).

## 260 **DUSCUSSION**

261 In this study, we assembled and validated the first two gap-free reference genome  
262 sequences of rice available to the research community. At present, this work could  
263 only have been achieved with a combination of multiple and deep-coverage sequence

264 datasets, cutting-edge technologies and assemblers, verses reliance on a single  
265 sequence dataset and assembler. For example, none of the *de novo* assemblers could  
266 ideally produce all complete pseudomolecules for the 12 rice chromosomes, but a set  
267 of fragmented contigs ([Supplemental Table 3](#)). Even with the same dataset, assembly  
268 results varied when using different assemblers and parameters. As we observed in our  
269 project, the data obtained by different sequencing approaches have different  
270 coverages: i.e. both the PacBio HiFi and CLR reads covered >99.9% of the ZS97RS3  
271 and MH63RS3 gap-free genomes, while BAC-based reads of RS1 assemblies only  
272 covered 88.59% and 90.95%, respectively ([Figure 1b](#)). Hence, the strategy applied  
273 here was to fully leverage the complementarity of datasets, assemblers and  
274 technologies. In our final assemblies, we manually selected and merged proper  
275 sequence contigs to cover their corresponding regions across each chromosome  
276 ([Supplemental Figure 6](#)). The last closed gaps in our assemblies were all in  
277 centromere regions, which confirms that the great challenge for completely  
278 assembling plant genomes is was from the nature of their complicated architecture  
279 and highly repetitive sequences. Long-read sequencing data of high accuracy,  
280 however, can span the repeats allowed assemblers to distinguish minor sequence  
281 differences in repeat regions during the assembling process.

282 We also used independent methods like Hi-C and Bionano maps to validate our  
283 assemblies, as well as FISH and ChIP-Seq assays to discover and characterize the  
284 location and primary structure of centromeres.

285 In conclusion, the generation and validation of two gap-free assemblies of ZS97  
286 and MH63, presented here, provides a clear picture of the primary sequence  
287 architecture of the *xian/indica* rice genomes that feed the world. Such data will serve  
288 as a fundamental and comprehensive model resource in the study of hybrid vigor, and  
289 other basic and applied research, and leads the path forward to a new standard for  
290 reference genomes in plant biology.

291

## 292 **METHODS**

### 293 **Plant Materials and Sequencing**

294 Fresh young leaf tissue was collected from *O. sativa* ZS97 and MH63 plants. We  
295 constructed SMRTbell libraries as described in previous study (Pendleton et al., 2015).  
296 The genomes of MH63 and ZS97 were sequenced using the PacBio Sequel II  
297 platform (Pacific Biosciences), to produce 8.34 Gb HiFi reads (~23x coverage) and  
298 48.39Gb CLR reads (~131x coverage) for ZS97, and 37.88 Gb HiFi reads (~103x  
299 coverage) and 48.97 Gb CLR reads (~132x coverage) for MH63 genomes.

300 Truseq Nano DNA HT Sample preparation Kit following manufacturer's standard  
301 protocol (Illumina) was used to generate the libraries for Illumina paired-end genome  
302 sequencing. These libraries were sequenced to generate 150 bp paired-end reads by  
303 Illumina HiSeq X Ten platform with 350 bp insert size, and produce 25 Gb reads  
304 (~69x coverage) for ZS97, and 28 Gb reads (~76x coverage) for MH63 genomes.

305 Plant tissues used for optical mapping were extracted using the Bionano plant  
306 tissue extraction protocol (Staňková et al., 2016). Extracted DNA was embedded in  
307 BioRad LE agarose for subsequent washes of T.E., proteinase K (0.8mg/ml), and  
308 RNase A (20µL/mL) treatments in lysis buffer. The agarose plugs were then melted  
309 using agarase (0.1 U/µL, New England Biolabs) and dialyzed on millipore  
310 membranes (0.1µm) with T.E. to equilibrate ion concentrations. The DNA was then  
311 nicked with the nickase restriction enzyme BssSI (2U/µL). Labeled nucleotides were  
312 incorporated at breakpoints and the DNA was counterstained. Each sample was  
313 loaded onto 2 nanochannel flow cells of a Bionano Irys machine for DNA imaging.

314

### 315 **Genome Assembly and Assessment**

316 Seven tools based on different algorithms were used to assemble the genomes of  
317 ZS97 and MH63. (1) Canu v1.8 (Koren et al., 2017) was used to assemble the  
318 genomes with default parameters; (2) FALCON toolkit v0.30 (Carvalho et al., 2016)  
319 was applied for assembly with the parameters: pa\_DBsplit\_option = -s200 -x500,

320 ovlp\_DBsplit\_option = -s200 -x500, pa\_REPmask\_code = 0,300;0,300;0,300,  
321 genome\_size = 400000000, seed\_coverage = 30, length\_cutoff = -1,  
322 pa\_HPCdaligner\_option = -v -B128 -M24, pa\_daligner\_option = -k18 -w8 -h480  
323 -e.80 -l5000 -s100, falcon\_sense\_option = --output-multi --min-idt 0.70 --min-cov 3  
324 --max-n-read 400, falcon\_sense\_greedy = False, ovlp\_HPCdaligner\_option = -v -M24  
325 -l500, ovlp\_daligner\_option = -h60 -e0.96 -s1000, overlap\_filtering\_setting =  
326 --max-diff 100 --max-cov 100- --min-cov 2, length\_cutoff\_pr = 1000; (3) MECAT2  
327 (Xiao et al., 2017) was utilized to assemble with the parameters: GENOME\_SIZE =  
328 400000000, MIN\_READ\_LENGTH = 2000, CNS\_OVLP\_OPTIONS = "",  
329 CNS\_OPTIONS = "-r 0.6 -a 1000 -c 4 -l 2000", CNS\_OUTPUT\_COVERAGE = 30,  
330 TRIM\_OVLP\_OPTIONS = "-B", ASM\_OVLP\_OPTIONS = "-n 100 -z 10 -b 2000 -e  
331 0.5 -j 1 -u 0 -a 400", FSA\_OL\_FILTER\_OPTIONS = "--max\_overhang = -1  
332 --min\_identity = - 1", FSA\_ASSEMBLE\_OPTIONS = "", GRID\_NODE = 0,  
333 CLEANUP = 0, USE\_GRID = false; (4) Flye 2.6-release (Kolmogorov et al., 2019)  
334 was set with "--genome-size 400m"; (5) Wtdbg2 2.5 (Ruan and Li., 2020) was used to  
335 assemble with parameters "-x sq, -g 400m", and then Minimap2 (Li, 2018) was  
336 employed to map the PacBio CLR data to the assembly results, and wtpoa was  
337 utilized to polish and correct the wtdbg2 assembly results; (6) NextDenovo v2.1-beta.0  
338 (<https://github.com/Nextomics/NextDenovo>) was applied for assembly with  
339 parameters "task = all, rewrite = yes, deltmp = yes, rerun = 3, input\_type = raw,  
340 read\_cutoff = 1k, seed\_cutoff = 44382, blocksize = 2g, pa\_correction = 20,  
341 seed\_cutfiles = 20, sort\_options = -m 20g -t 10 -k 40, minimap2\_options\_raw = -x  
342 ava-ont -t 8, correction\_options = -p 10, random\_round = 20, minimap2\_options\_cns  
343 = -x ava-pb -t 8 -k17- w17, nextgraph\_options = -a 1"; (7) Miniasm-0.3-r179 (Li,  
344 2016) with default parameters.

345 Based on the results of these seven software tools, Genome Puzzle Master (GPM)  
346 (Zhang et al., 2016b) was then used to integrate and optimize the assembled contigs,  
347 and visualize complete chromosomes. Based on the HiFi and CLR sequencing data,

348 we used the GenomicConsensus package of SMRTLink/7.0.1.66975  
349 (<https://www.pacb.com/support/>) to polish the assembled genomes twice with the  
350 Arrow algorithm, using the parameter: --algorithm=arrow. Pilon (Walker et al., 2014)  
351 was used for polishing the genomes based on Illumina data with the parameters: --fix  
352 snps, indels. This process was repeated twice. Molecules were then assembled using  
353 Bionano IrysSolve pipeline (<https://bionanogenomics.com/support-page/>) to create  
354 optical maps. Images were interpreted quantitatively using Bionano AutoDetect  
355 2.1.4.9159 and data was visualized using IrysView v2.5.1. These assemblies were  
356 used with draft genome assemblies to validate and scaffold the sequences. Bionano  
357 optical map data was aligned to the merged contigs using RefAlignerAssembler in the  
358 IrysView software package to perform the verification.

359 ZS97RS3 and MH63RS3 genome completeness was assessed using BUSCO  
360 v4.0.6, which contained 1614 genes in the ‘embryophyta\_odb10’ dataset (Simão et al.,  
361 2015), with default parameters. In addition, we mapped the PacBio HiFi reads and  
362 PacBio CLR reads with Minimap2 (Li, 2018), Illumina reads with BWA-0.7.17 (Jo et  
363 al., 2015), BES/BAC reads with BLASTN v2.7.1 (Altschul et al., 1990), Hi-C reads  
364 with HiC-Pro v2.11.1 (Servant et al., 2015), RNA-Seq reads with Hisat2 v2.1.0 (Kim  
365 et al., 2015) to both genome assemblies.

366

### 367 **Gene and Repeat Annotations**

368 MAKER-P (Campbell et al., 2014) version 3 was used to annotate the ZS97RS3 and  
369 MH63RS3 genomes. All evidence was the same as that used for RS1 genome  
370 annotations. To ensure the consistency with the RS1 versions, genes that mapped in  
371 the entirety to the RS3 genomes were retained. New genes in gap regions were  
372 obtained from MAKER-P results (Campbell et al., 2014). Genes encoding  
373 transposable elements were identified and transitively annotated by searching against  
374 the MIPS-REdat Poaceae version 9.3p (Nussbaumer et al., 2013) database using  
375 TBLASTN (Altschul et al., 1990) with an E-value of 1e-10. tRNAs were identified

376 with tRNAscan-SE (Lowe and Eddy, 1997) using default parameters; rRNA genes  
377 were identified by searching each assembly against the rRNA sequences of  
378 Nipponbare using BLASTN v2.7.1 (Altschul et al., 1990); miRNAs and snRNAs  
379 were predicted using INFERNAL of Rfam (Griffiths-Jones et al., 2005) (v14.1).  
380 Repeats in the genome were annotated using RepeatMasker (Zhi et al., 2006) with  
381 RepBase (Bao et al., 2015), and TIGR Oryza Repeats (v3.3) with RMBlast search  
382 engine. For the overlapping repeats in different classes, LTR retrotransposons were  
383 kept first, next TIR, and then SINE and LINE, finally helitrons. This priority order  
384 was based on stronger structural signatures. Besides, the known nested insertions  
385 models (LTR into helitron, helitron into LTR, TIR into LTR, LTR into TIR) were  
386 retained. The identified repetitive elements were further characterized and classified  
387 using PGSB repeat classification schema. LTR\_FINDER (Xu and Wang 2007) was  
388 used to identify complete LTR-RTs with target site duplications (TSDs), primer  
389 binding sites (PBS) and polypurine tract (PPT).

390

### 391 **Chromatin Immunoprecipitation (ChIP) and ChIP-seq**

392 Procedures for chromatin immunoprecipitation (ChIP) were adopted from Nagaki et al.  
393 (2003) and Walkowiak et al. (2020) using nuclei from 4-week-old seedlings.  
394 Chromatin with the nuclei was digested with micrococcal nuclease (Sigma-Aldrich, St.  
395 Louis, MO) to liberate nucleosomes. For ChIP, we used anti-centromeric histone 3  
396 antibody (N-term) which reacts with 18.5 kDa CenH3 protein from *Oryza sativa*  
397 purchased from Antibodies-online Inc. (Limerick, PA; cat# ABIN1106669). The  
398 digested mixture was then incubated overnight with 3 µg of rice CENH3 antibody at  
399 4°C. The target antibodies were then captured from the mixture using Dynabeads  
400 Protein G (Invitrogen, Carlsbad, CA). ChIP-seq libraries were then constructed using a  
401 TruSeq ChIP Library Preparation Kit (Illumina, San Diego, CA) following the  
402 manufacturer's instructions and the libraries were sequenced (2x150bp) on an Illumina  
403 HiSeqX10.

404

## 405 **Fluorescence *in situ* Hybridization (FISH)**

### 406 *Slide Preparation*

407 Mitotic chromosomes were prepared as described by [Koo and Jiang \(2009\)](#) with  
408 minor modifications. Root tips were collected from plants and treated in a nitrous oxide  
409 gas chamber for 1.5 h. The root tips were fixed overnight in ethanol:glacial acetic acid  
410 (3:1) and then squashed in a drop of 45% acetic acid.

### 411 *Probe Labeling and Detection*

412 The ChIPed DNAs were labeled with digoxigenin-16-dUTP using a nick translation  
413 reaction. The clone, maize 45S rDNA ([Koo and Jiang 2009](#)) was labeled with  
414 biotin-11-dUTP (Roche, Indianapolis, IN). Biotin- and digoxigenin-labeled probes  
415 were detected with Alexa Fluor 488 streptavidin antibody (Invitrogen) and  
416 rhodamine-conjugated anti-digoxigenin antibody (Roche), respectively.  
417 Chromosomes were counterstained with 4',6-diamidino-2-phenylindole (DAPI) in  
418 Vectashield antifade solution (Vector Laboratories, Burlingame, CA). Images were  
419 captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy LLC,  
420 Thornwood, NY) using a cooled CCD camera CoolSNAP HQ2 (Photometrics,  
421 Tucson, AZ) and AxioVision 4.8 software. The final contrast of the images was  
422 processed using Adobe Photoshop CS5 software.

423

## 424 **The Completeness of Centromeres on MH63RS3 and ZS97RS3 Chromosomes**

425 Based on the final RS3 genome assemblies, we use BLAST ([Altschul et al., 1990](#)) to  
426 align the *CentO* satellite repeats in rice to each reference genome with an E-value of  
427  $1e-5$ , and then use BEDtools ([Quinlan, 2014](#)) to merge the results with the parameter  
428 '-d 50000'. If a region contained more than 10 consecutive *CentO* repeats with  
429 lengths longer than 10 kb, it was classified as core centromere region (CCR).

430 For the identification of the whole centromere region, we use BWA-0.7.17 ([Jo  
431 and Koh., 2015](#)) to align the CENH3 ChIP-Seq reads to MH63RS3 and ZS97RS3  
432 genomes, and use SAMtools ([Li et al., 2009](#)) to filter the results with mapQ value



433 above 30; then we use MACS2 (Zhang et al., 2008) to call the peaks of CENH3.  
434 Finally, we manually defined all the centromeric region of MH63RS3 and ZS97RS3  
435 genomes in consideration of the distribution of CENH3 histone, *CentO*, repeats and  
436 genes.

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

443

#### 444 **Telomere Sequence Identification**

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

449

#### 450 **Identification of PAVs between ZS97RS3 and MH63RS3**

451 ZS97RS3 assembly was aligned to MH63RS3 assembly using Mummer (4.0.0beta2)  
452 (Marçais et al., 2018) with parameters settings '-c 90 -l 40'. Then we used "delta-filter  
453 -1" parameter with the one-to-one alignment block option to filter the alignment  
454 results. Further "show-diff" was used to select for unaligned regions as the PAVs.

455

#### 456 **Prediction of NLR Genes**

457 We first predicted domains of genes with InterProScan (Jones et al., 2014), which can  
458 analyze peptide sequences against InterPro member databases, including ProDom,  
459 PROSITE, PRINTS, Pfam, PANTHER, SMART and Coils. Pfam and Coils were used  
460 to prediction NLRs which were required to contain at least one NB, TIR, or CC<sub>R</sub>

461 (RPW8) using the following reference sequences: NB (Pfam accession PF00931),  
462 TIR (PF01582), RPW8 (PF05659), LRR (PF00560, PF07725, PF13306, PF13855)  
463 domains, or CC motifs ([Van de Weyer et al., 2019](#)).

464

#### 465 **Identification of Collinear Orthologues**

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

469

#### 470 **DATA AVAILABILITY**

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

#### 477 **FUNDING**

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

#### 483 **AUTHOR CONTRIBUTIONS**

484 L.-L.C., J.Z., R.W. and Q.Z. designed studies and contributed to the original concept  
485 of the project. J.P. and D.-H.K. performed the ChIP-seq and FISH experiments. D.K.,  
486 E.L., S.L., J.T., D.Y., J.U. and R.W. performed the genome and Bionano sequencing.  
487 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

488 genome assembling and annotation, comparative genomics analysis and other data  
489 analysis. J.-M.S., W.-Z.X., S.W., J.P., D.-H.K., L.-L.C. and J.Z. wrote the paper.  
490 W.X., R.W. and Q.Z. contributed to revisions.

491

## 492 **ACKNOWLEDGEMENTS**

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

497

## 498 **ONLINE CONTENT**

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

## 502 **REFERENCES**

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

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

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

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

513 **Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R.,**

- 514 **and Jiang, J.** (2002). Functional rice centromeres are marked by a satellite repeat  
515 and a centromere-specific retrotransposon. *Plant Cell* **14**:1691–1704.
- 516 **Simão, F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., and Zdobnov,**  
517 **E.M.** (2015). BUSCO: assessing genome assembly and annotation completeness  
518 with single-copy orthologs. *Bioinformatics* **31**:3210–3212.
- 519 **Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., and Zhang, Q.** (2006).  
520 GS3, a major QTL for grain length and weight and minor QTL for grain width  
521 and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl.*  
522 *Genet.* **112**:1164–1171.
- 523 **Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W.** (2012). CD-HIT: accelerated for  
524 clustering the next generation sequencing data. *Bioinformatics* **28**:3150–3152.
- 525 **Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and**  
526 **Bateman, A.** (2005). Rfam: annotating non-coding RNAs in complete genomes.  
527 *Nucleic Acids Res.* **33**:D121–D124.
- 528 **Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S., and Zhang, Q.** (2003).  
529 Single-locus heterotic effects and dominance by dominance interactions can  
530 adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc.*  
531 *Natl. Acad. Sci. USA* **100**:2574–2579.
- 532 **Hua, J.P., Xing, Y.Z., Xu, C.G., Sun, X.L., Yu, S.B., and Zhang, Q.** (2002). Genetic  
533 dissection of an elite rice hybrid revealed that heterozygotes are not always  
534 advantageous for performance. *Genetics* **162**: 885–1895.
- 535 **Huang, Y., Zhang, L., Zhang, J., Yuan, D., Xu, C., Li, X., Zhou, D., Wang, S., and**  
536 **Zhang, Q.** (2006). Heterosis and polymorphisms of gene expression in an elite  
537 rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. *Plant Mol.*  
538 *Biol.* **62**:579–591.
- 539 **Jo, H., and Koh, G.** (2015). Faster single-end alignment generation utilizing  
540 multi-thread for BWA. *Biomed Mater Eng. Suppl* **1**:S1791-1796.
- 541 **Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam,**  
542 **H., Maslen, J., Mitchell, A., Nuka, G., et al.** (2014). InterProScan 5:  
543 genome-scale protein function classification. *Bioinformatics* **30**:1236–1240.

- 544 **Kent, W.J.** (2002). BLAT--the BLAST-like alignment tool. *Genome Res.* **12**:656-664.
- 545 **Kim, D., Langmead, B., and Salzberg, S.L.** (2015). HISAT: a fast spliced aligner  
546 with low memory requirements. *Nat. Methods* **12**:357–360.
- 547 **Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A.** (2019). Assembly of long,  
548 error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**:540–546.
- 549 **Koo, D.H., and Jiang, J.M.** (2009). Super-stretched pachytene chromosomes for  
550 fluorescence in situ hybridization mapping and immunodetection of cytosine  
551 methylation. *Plant J.* **59**:509–516.
- 552 **Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy,**  
553 **A.M.** (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer  
554 weighting and repeat separation. *Genome Res.* **27**:722–736.
- 555 **Li, H.** (2016). Minimap and minimap: fast mapping and de novo assembly for noisy  
556 long sequences. *Bioinformatics* **32**:2103–2110.
- 557 **Li, H.** (2018). Minimap2: pairwise alignment for nucleotide sequences.  
558 *Bioinformatics* **34**:3094–3100.
- 559 **Li, H., and Durbin, R.** (2009). Fast and accurate short read alignment with  
560 Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.
- 561 **Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,**  
562 **Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing**  
563 **Supgroup.** (2009). The Sequence Alignment/Map format and SAMtools.  
564 *Bioinformatics* **25**:2078–2079.
- 565 **Logsdon, G.A., Vollger, M.R., Hsieh, P.H., Mao, Y., Liskovych, M.A., Koren, S.,**  
566 **Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al.** (2020). The structure,  
567 function, and evolution of a complete human chromosome 8. *bioRxiv*  
568 2020.09.08.285395
- 569 **Lowe, T.M., and Eddy, S.R.** (1997). tRNAscan-SE: a program for improved  
570 detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*  
571 **25**:955–964.

- 572 **Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and**  
573 **Zimin, A.** (2018). MUMmer4: A fast and versatile genome alignment system.  
574 *PLoS Comput. Biol.* **14**:e1005944.
- 575 **Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A.,**  
576 **Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al.** (2020).  
577 Telomere-to-telomere assembly of a complete human X chromosome. *Nature*  
578 **585**:79–84.
- 579 **Mussurova, S., Al-Bader, N., Zuccolo, A., and Wing, R.A.** (2020). Potential of  
580 platinum standard reference genomes to exploit natural variation in the wild  
581 relatives of rice. *Front Plant Sci.* **11**:579980.
- 582 **Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., and Jiang, J.**  
583 (2003). Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is  
584 the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics*  
585 **163**:1221–1225.
- 586 **Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma,**  
587 **S., Gundlach, H., and Spannagl, M.** (2013). MIPS PlantsDB: a database  
588 framework for comparative plant genome research. *Nucleic Acids Res.*  
589 **41**:D1144-1151.
- 590 **Ou, S., Chen, J., and Jiang, N.** (2018). Assessing genome assembly quality using the  
591 LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**:e126.
- 592 **Ouyang, S., and Buell, C.R.** (2004). The TIGR Plant Repeat Databases: a collective  
593 resource for the identification of repetitive sequences in plants. *Nucleic Acids*  
594 *Res.* **32**:D360–363.
- 595 **Pierce, N.T., Irber, L., Reiter, T., Brooks, P., and Brown, C.T.** (2019). Large-scale  
596 sequence comparisons with *sourmash*. *F1000Res* **8**:1006.
- 597 **Pendleton, M., Sebra, R., Pang, A.W., Ummat, A., Franzen, O., Rausch, T., Stütz,**  
598 **A.M., Stedman, W., Anantharaman, T., Hastie, A., et al.** (2015). Assembly and  
599 diploid architecture of an individual human genome via single-molecule  
600 technologies. *Nat. Methods* **12**:780-786.

- 601 **Perumal, S., Koh, C.S., Jin, L., Buchwaldt, M., Higgins, E.E., Zheng, C., Sankoff,**  
602 **D., Robinson, S.J., Kagale, S., Navabi, Z.K., et al.** (2020). A high-contiguity  
603 *Brassica nigra* genome localizes active centromeres and defines the ancestral  
604 Brassica genome. *Nat Plants* **6**:929-941.
- 605 **Quinlan, A.R.** (2014). BEDTools: the swiss-army tool for genome feature analysis.  
606 *Curr. Protoc. Bioinformatics* **47**:11.12.134.
- 607 **Rice Chromosomes 11 and 12 Sequencing Consortia.** (2005). The sequence of rice  
608 chromosomes 11 and 12, rich in disease resistance genes and recent gene  
609 duplications. *BMC Biol.* **3**:20.
- 610 **Ruan, J., and Li, H.** (2020). Fast and accurate long-read assembly with wtdbg2. *Nat.*  
611 *Methods* **17**:155–158.
- 612 **Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard,**  
613 **E., Dekker, J., and Barillot, E.** (2015). HiC-Pro: an optimized and flexible  
614 pipeline for Hi-C data processing. *Genome Biol.* **16**:259.
- 615 **Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.**  
616 (2015). BUSCO: assessing genome assembly and annotation completeness  
617 with single-copy orthologs. *Bioinformatics* **31**:3210-3212.
- 618 **Staňková, H., Hastie, A.R., Chan, S., Vrána, J., Tulpová, Z., Kubaláková, M.,**  
619 **Visendi, P., Hayashi, S., Luo, M., Batley, J., et al.** (2016). BioNano genome  
620 mapping of individual chromosomes supports physical mapping and sequence  
621 assembly in complex plant genomes. *Plant Biotechnol J.* **14**:1523-1531.
- 622 **Sun, X., Cao, Y., Yang, Z., Xu, C., Li, X., Wang, S., and Zhang, Q.** (2004) Xa26, a  
623 gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes  
624 an LRR receptor kinase-like protein. *Plant J.* **37**:517–527.
- 625 **Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H.** (2008).  
626 Synteny and collinearity in plant genomes. *Science* **320**:486–488.
- 627 **Van de Weyer, A.-L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V.,**  
628 **Witek, K., Jones, J.D.G., Dangl, J.L., Weigel, D., and Bemm, F.** (2019). A

629 species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell*  
630 **178**:1260–1272.

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

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

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

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

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

648 **Xu, Z., and Wang, H.** (2007). LTR\_FINDER: an efficient tool for the prediction of  
649 full-length LTR retrotransposons. *Nucleic Acids Res.* **35**:W265-268.

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

653 **Zhang, J., Chen, L.L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li,**  
654 **W., Song, J.M., Xie, W., et al.** (2016a). Extensive sequence divergence between  
655 the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui  
656 63. *Proc. Natl. Acad. Sci. USA* **113**:E5163–5171.

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



660           Bioinformatics **32**:3058–3064.

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

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

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

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

673

674

675

676

677

678

679

680

681

682 **FIGURE LEGENDS**

683 **Figure 1. Two gap-free genomes of rice.**

684 (A) Collinearity analysis between ZS97RS3 and MH63RS3. The collinear regions  
685 between ZS97RS3 and MH63RS3 are shown linked by gray lines. All the RS1 gap  
686 regions closed in RS3 are showed in yellow blocks. The black triangle indicates  
687 presence of telomere sequence repeats. Repeat percentage distribution is plotted  
688 above/under each chromosome in 100-kb bins; (B) Histogram showed the reads  
689 coverage for different libraries in MH63RS3 and ZS97RS3, including BAC, CCS and  
690 CLR reads.

691 **Figure 2. Structural variations of ZS97RS3 and MH63RS3.**

692 (A) Distribution of the difference regions between ZS97RS3 and MH63RS3 on the  
693 chromosome. (B) Distribution of the NLR genes of ZS97RS3 and MH63RS3 on the  
694 chromosome. (C) The expansion structural variation MH-E in MH63RS3. The  
695 structural of MH-E at the end of chromosome 11 of MH63RS3, from top to bottom  
696 are the gene collinearity of ZS97RS3 and MH63RS3, the TE distribution, the gene  
697 expression in this region. (D) The insertion structural variation MH-I in MH63RS3.  
698 From top to bottom are the gene collinearity of ZS97RS3 and MH63RS3, the TE  
699 distribution and the gene expression in this region. (E) Coverage ratio of two  
700 structural variations (MH-E and MH-I) in 25 rice varieties.

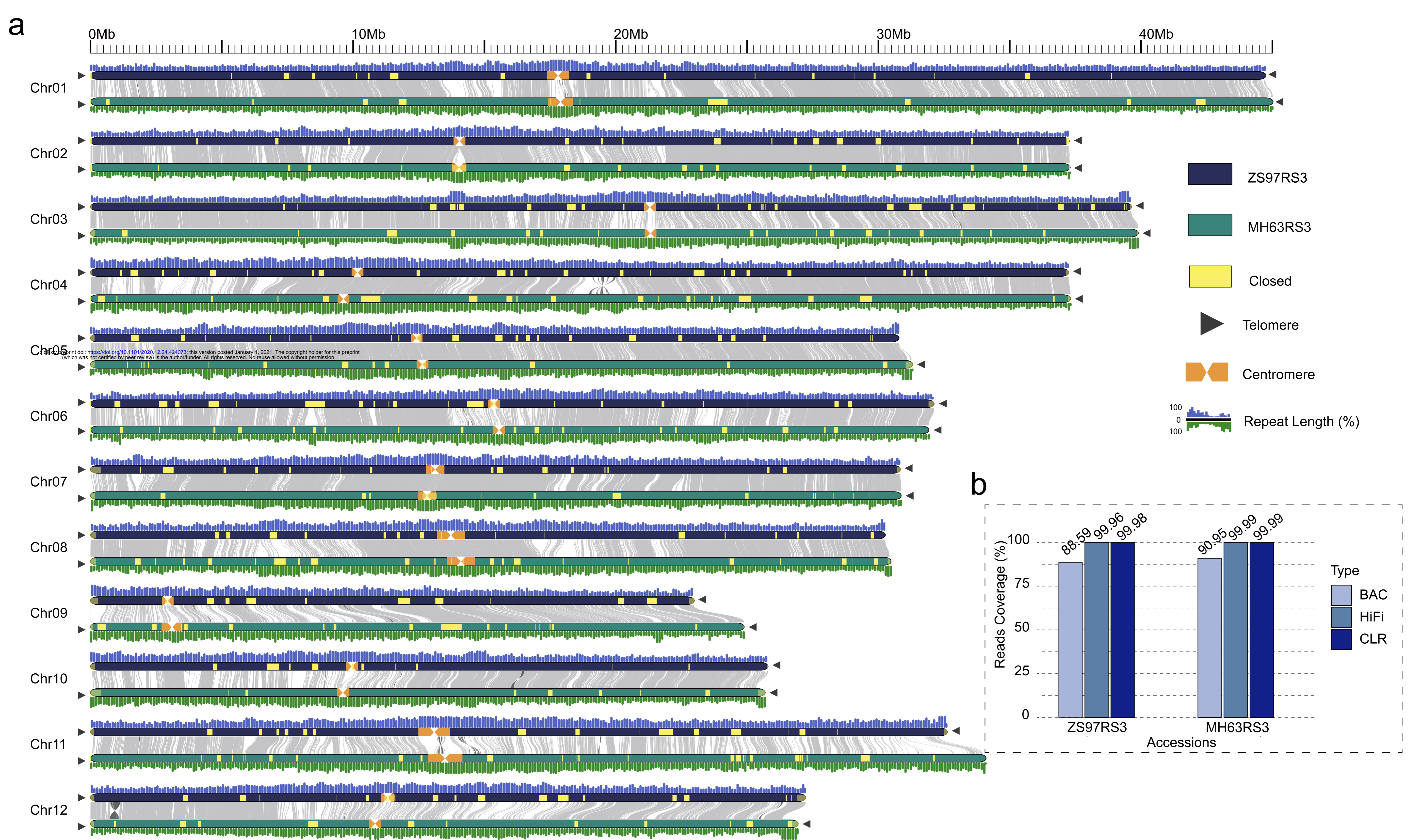
701 **Figure 3. Characterization of complete rice centromeres.**

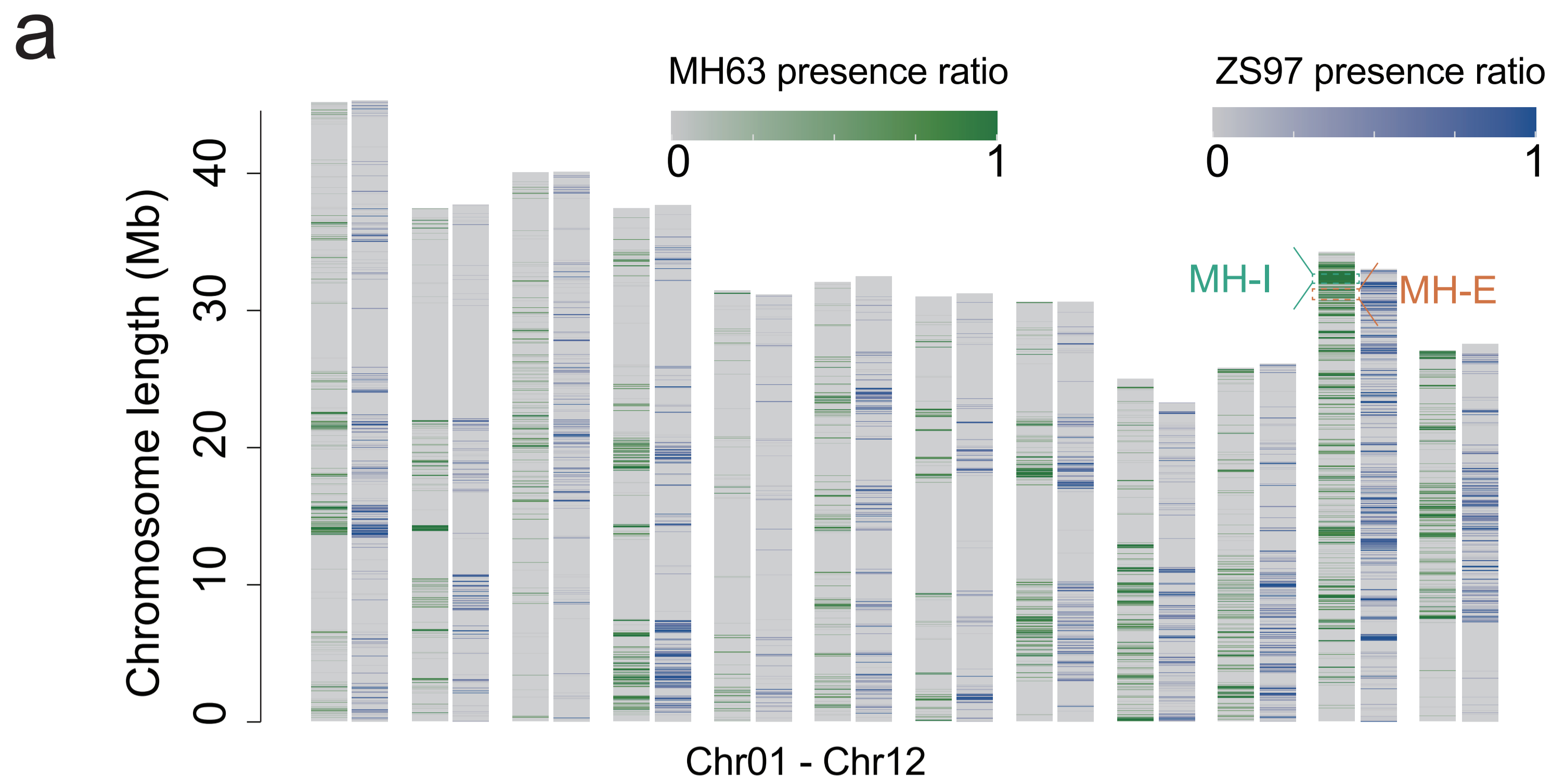
702 (A) The definition of MH63RS3 centromere. The layers of each chromosome graph  
703 indicate 1) the density of read mapping from CENH3 Chip-seq with sliding windows  
704 of 10-kb and 20-kb shown in grey and blue lines respectively, 2) the *CentO* satellite  
705 distribution, 3) non-TE genes distribution, and 4) TE distribution, respectively. The  
706 dotted frame represents the defined centromere area. (B) Fluorescence *in situ*  
707 hybridization (FISH) of mitotic metaphase chromosomes in MH63 and ZS97 using

708 CENH3 ChIP-DNA as probe (red) with chromosomes counterstained with DAPI  
709 (blue). **(C)** Coverage of HiFi, CLR, Illumina reads and distribution of TEs in the  
710 centromere on Chr01 (extended 500 kb left and right) of MH63RS3. **(D)** The pairwise  
711 synteny visualization between ZS97RS3 and MH63RS3 in centromere area of Chr01.  
712 The synteny genes between ZS97RS3 and MH63RS3 were linked as the gray lines.  
713 The yellow blocks were core regions. **(E)** Characteristics of the centromere on Chr01  
714 of MH63RS3. The first layer is histone CENH3 distribution, the second layer is the  
715 *CentO* distribution, the third layer is the Genes distribution, the fourth to sixth levels  
716 are gene expression, the seventh to ninth levels are methylation distribution, the tenth  
717 layer is *CentO* sequence similarity.

718

719





bioRxiv preprint doi: <https://doi.org/10.1101/2020.12.24.424073>; this version posted January 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

