

1 **Twelve Platinum-Standard Reference Genomes Sequences (PSRefSeq) that**
2 **complete the full range of genetic diversity of Asian rice**

3
4 Yong Zhou^{1a}, Dmytro Chebotarov^{2a}, Dave Kudrna³, Victor Llaca⁴, Seunghee Lee³,
5 Shanmugam Rajasekar³, Nahed Mohammed¹, Noor Al-Bader¹, Chandler Sobel-
6 Sorenson³, Praveena Parakkal⁴, Lady Johanna Arbelaez⁵, Natalia Franco⁵, Nickolai
7 Alexandrov², N. Ruaraidh Sackville Hamilton², Hei Leung², Ramil Mauleon², Mathias
8 Lorieux^{5,6}, Andrea Zuccolo^{1,7*}, Kenneth McNally^{2*}, Jianwei Zhang^{3,8*}, Rod A. Wing^{1,2,3*}

9
10 ¹Center for Desert Agriculture, Biological and Environmental Sciences & Engineering
11 Division (BESE), King Abdullah University of Science and Technology (KAUST),
12 Thuwal, 23955-6900, Saudi Arabia

13 ²International Rice Research Institute (IRRI), Strategic Innovation, Los Baños, 4031
14 Laguna, Philippines

15 ³Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson,
16 Arizona 85721, USA

17 ⁴Genomics Technologies, Applied Science and Technology, Corteva Agriscience™, IA
18 50131, USA

19 ⁵Rice Genetics and Genomics Lab, International Center for Tropical Agriculture (CIAT),
20 Cali, Colombia

21 ⁶University of Montpellier, DIADE, IRD, France

22 ⁷Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy

23 ⁸National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural
24 University, Wuhan 430070, China

25 Yong Zhou^{1a}, yong.zhou@kaust.edu.sa, ORCID 0000-0002-1662-9589

26 Dmytro Chebotarov^{2a}, d.chebotarov@irri.org, ORCID 0000-0003-1351-9453

27 Dave Kudrna³, dkudrna@email.arizona.edu, ORCID 0000-0002-3092-3629

28 Victor Llaca⁴, victor.llaca@corteva.com, ORCID 0000-0003-4822-2924

29 Seunghee Lee³, seunghl@ag.arizona.edu,

30 Shanmugam Rajasekar³, shans@email.arizona.edu,

31 Nahed Mohammed¹, nahed.mohammed@kaust.edu.sa, ORCID 0000-0002-8857-3246

32 Noor Al-Bader¹, noor.albader@kaust.edu.sa, ORCID 0000-0002-0511-6972

33 Chandler Sobel-Sorenson³, scar@email.arizona.edu,

34 Praveena Parakkal⁴, Praveena.parakkal@corteva.com,

35 Lady Johanna Arbelaez⁵, layoar@gmail.com,
36 Natalia Franco⁵, n.franco@cgiar.org,
37 Nickolai Alexandrov², nickolai.alexandrov@gmail.com, ORCID 0000-0003-3381-0918
38 N. Ruairaidh Sackville Hamilton², ruairaidh.sh@gmail.com>, ORCID 0000-0002-8467-
39 0110
40 Hei Leung², h.leung@irri.org,
41 Ramil Mauleon², ramil.mauleon@scu.edu.au, ORCID 0000-0001-8512-144X
42 Current address: Southern Cross Plant Science, Southern Cross University, Lismore,
43 Australia
44 Mathias Lorieux^{5,6}, mathias.lorieux@ird.fr, ORCID 0000-0001-9864-3933
45 Andrea Zuccolo^{1,7*}, andrea.zuccolo@kaust.edu, ORCID 0000-0001-7574-0714
46 Kenneth McNally^{2*}, k.mcnally@irri.org, ORCID 0000-0002-9613-5537
47 Jianwei Zhang^{3,8*}, jzhang@mail.hzau.edu.cn, ORCID 0000-0001-8030-5346
48 Rod A. Wing^{1,2,3*}, rwing@email.arizona.edu, ORCID 0000-0001-6633-622
49

50 ^aThese authors contributed equally to this work.

51 *Correspondence and requests for materials should be addressed to: Andrea Zuccolo
52 (email: andrea.zuccolo@kaust.edu), Kenneth McNally (email: k.mcnally@irri.org),
53 Jianwei Zhang (email: jzhang@mail.hzau.edu.cn), or Rod A. Wing (email:
54 rod.wing@kaust.edu.sa; rwing@email.arizona.edu).

55 **Abstract**

56

57 As the human population grows from 7.8 billion to 10 billion over the next 30 years,
58 breeders must do everything possible to create crops that are highly productive and
59 nutritious, while simultaneously having less of an environmental footprint. Rice will play
60 a critical role in meeting this demand and thus, knowledge of the full repertoire of genetic
61 diversity that exists in germplasm banks across the globe is required. To meet this
62 demand, we describe the generation, validation and preliminary analyses of transposable
63 element and long-range structural variation content of 12 near-gap-free reference genome
64 sequences (RefSeqs) from representatives of 12 of 15 subpopulations of cultivated rice.
65 When combined with 4 existing RefSeqs, that represent the 3 remaining rice
66 subpopulations and the largest admixed population, this collection of 16 Platinum
67 Standard RefSeqs (PSRefSeq) can be used as a pan-genome template to map
68 resequencing data to detect virtually all standing natural variation that exists in the pan-
69 cultivated rice genome.

70 **Background & Summary**

71 Asian cultivated rice is a staple food for half of the world population. With the planet's
72 population expected to reach 10 billion by 2050, farmers must increase production by at
73 least 100 million metric tons per year (Seck et al 2012; Merrey et al. 2018). To address
74 this need, future rice cultivars should provide higher yields, be more nutritious, be
75 resilient to multiple abiotic and biotic stresses, and have less of an environmental
76 footprint (Wing et al. 2018; 3K RGP 2014). To achieve this goal, a comprehensive and
77 more in-depth understanding of the full range of genetic diversity of the pan-cultivated
78 rice genome and its wild relatives will be needed (Stein et al. 2018).

79 With a genome size of ~390 Mb, rice has the smallest genome among the
80 domesticated cereals, making it particularly amenable to genomic studies (Kawahara et
81 al. 2013) and the primary reason why it was the first crop genome to be sequenced 15
82 years ago (International Rice Genome Sequencing 2005). To better understand the full-
83 range of genetic diversity that is stored in rice germplasm banks around the world,
84 several studies have been conducted using microarrays (Thomson et al. 2017; McNally et
85 al. 2009) and low coverage skim sequencing (Huang et al. 2012; Zhao et al. 2018). In
86 2018, a detailed analysis of the Illumina resequencing of more than 3,000 diverse rice
87 accessions (a.k.a. 3K-RG), aligned to the *O. sativa* v.g. japonica cv. Nipponbare
88 reference genome sequence (a.k.a. IRGSP RefSeq), showed how the high genetic
89 diversity present in domesticated rice populations provides a solid base for the
90 improvement of rice cultivars (Wang et al. 2018). One key finding from a population
91 structure analysis of this dataset showed that the 3,000 accessions can be subdivided into
92 nine subpopulations, where most accessions from close sub-groups could be associated to
93 geographic origin (Wang et al. 2018).

94 One critical piece of information missing from these analyses is the fact that single
95 nucleotide polymorphisms (SNPs) and structural variations (SVs) present in
96 subpopulation specific genomic regions have yet to be detected because the 3K-RG data
97 set was only aligned to a single reference genome. Therefore, the next logical step, to
98 capture and understand genetic variation pan-subpopulation-wide is to map the 3K-RG
99 dataset to high-quality reference genomes that represent each of the subpopulations of
100 cultivated Asian rice. At present, only a handful high-quality rice genomes for cultivated
101 rice are publicly available (Kawahara et al. 2013, Zhang et al. 2016a, Zhang et al. 2016b
102 and Stein et al. 2018), thus, there is an immediate need for such a comprehensive
103 resource to be created, which is the subject of this Data Descriptor.

104 Here we present a reanalysis of the population structure analysis discussed above
105 (Wang et al. 2018) and show that the 3K-RG dataset can be further subdivided into a total
106 of 15 subpopulations. We then present the generation of 12 new and near-gap-free high-
107 quality PacBio long-read reference genomes from representative accessions of the 12
108 subpopulations of cultivated rice for which no high-quality reference genomes exist. All
109 12 genomes were assembled with more than 100x genome coverage PacBio long-read
110 sequence data and then validated with Bionano optical maps (Udall and Dawe 2018). The
111 number of contigs covering each of the twelve 12 assemblies, excluding unplaced
112 contigs, ranged from 15 (GOBOL SAIL (BALAM)::IRGC 26624-2) to 104 (IR 64). The
113 contig N50 value for the 12 genome data set ranged from 7.35 Mb to 31.91 Mb. When
114 combined with 4 previously published genomes (i.e. Minghui 63 (MH 63), Zhenshan 97
115 (ZS 97) (Zhang et al. 2016a, b), N 22 (Stein et al. 2018; updated in 2019) and the IRGSP
116 RefSeq (Kawahara et al. 2013)), this 16 genome dataset can be used to represent the
117 K=15 population/admixture structure of cultivated Asian rice.

118 **Methods**

119 **Ethics statement**

120 This work was approved by the University of Arizona (UA), the King Abdullah
121 University of Science and Technology (KAUST), Huazhong Agricultural University
122 (HZAU), the International Rice Research Institute (IRRI) and the International Center for
123 Tropical Agriculture (CIAT). All methods used in this study were carried out following
124 approved guidelines.

125 **Population structure**

126 We extracted 30 subsets of 100,000 randomly chosen SNPs out of the 3K-RG Core SNP
127 set v4 (996,009 SNPs, available at https://snp-seek.irri.org/_download.zul). For each
128 subset, we ran ADMIXTURE (Alexander et al. 2009) with the number of ancestral
129 groups K ranging from 5 to 15. We then aligned the resulting Q matrices using CLUMPP
130 software (Jakobsson and Rosenberg 2007). Since different runs at a given value of K
131 often give rise to different refinements (splits) of the lower level grouping, we first
132 clustered the runs for each K according to similarity of Q matrices using hierarchical
133 clustering, thus obtaining several clusters of runs for each K. We discarded one-element
134 clusters (outlier runs), and averaged the Q matrices within each remaining cluster. Figure
135 S1 shows the admixture proportions taken from the averaged Q matrices of the final

136 clusters for K=9 to 15. The columns of these averaged Q matrices, representing
137 admixture proportions for groups discovered in different runs, were then used to define
138 the “K15” grouping. At K=9, 12, and 13, the Q matrices converged to two different
139 modes according to whether XI-1A or GJ-trop is split (these are labeled as K=9.1, 12.1
140 and 13.1).

141 The group membership for each sample was defined by applying the threshold of 0.65
142 to admixture components. Samples with no admixture components exceeding 0.65 were
143 classified as follows. If the sum of components for subpopulations within the major
144 groups cA (*circum*-Aus), XI (*Xian*-indica), and GJ (*Geng*-japonica) was ≥ 0.65 , the
145 samples were classified as cA-adm (admixed within cA), XI-adm (within XI) or GJ-adm
146 (within GJ), respectively, and the remaining samples were deemed ‘fully’ admixed.
147 The newly defined groups were mostly either aligned with the previous K=9 grouping, or
148 refined those groups, and they were named accordingly (e.g. XI-1B1 and XI-1B2 are new
149 subgroups within XI-1B).

150 The phenogram shown in Figure 1 was constructed with DARwin v6
151 (<http://darwin.cirad.fr/>, unweighted Neighbor-joining) using the identity by state (IBS)
152 distance matrix from Plink on the 4.8M Filtered SNP set (available at [https://snp-
153 seek.irri.org/_download.zul](https://snp-seek.irri.org/_download.zul)). Colors were assigned to subpopulations based on K15
154 Admixture results. One entry, MH 63 (XI-adm) represents the admixed types among the
155 XI group.

156 **Sample selection, collection and nucleic acid preparation**

157 To select accessions to represent the 12 subpopulations of Asian rice that lack high-
158 quality reference genome assemblies, the following strategy was employed. The IBS
159 distance matrix was used for a principal component analysis (PCA) analysis in R to
160 generate 5 component axes. Then, for each of the 12 subpopulations, i.e. *circum*-Aus2 =
161 cA2, *circum*-Basmati = cB, *Geng*-japonica (GJ) subtropical (GJ-subtrp), tropical1 (GJ-
162 trop1) and tropical2 (GJ-trop2), and *Xian*-indica (XI) subpopulations XI-1B1, XI-1B2,
163 XI-2A, XI-2B, XI-3A, XI-3B1, XI-3B2, the centroid of each group in the space spanned
164 by first 5 principal components was determined from the eigenvectors, and the entry
165 closest to the centroid for which seed was available was chosen as the representative for
166 that subpopulation (Table 1).

167 Single seed decent (SSD) seed from IR 64 and Azucena were obtained from the Rice
168 Genetics and Genomics Laboratory, CIAT, in Cali, Colombia, and SSD seed from the

169 remaining 10 accessions (Table 1) were obtained from the International Rice Genebank,
170 maintained by IRRI, Los Baños, Philippines. All seed were sown in potting soil and
171 grown under standard greenhouse conditions at UA, Tucson, USA for 6 weeks at which
172 point they were dark treated for 48-hours to reduce starch accumulation. Approximately
173 20-50 grams of young leaf tissue was then harvested from each accession and
174 immediately flash frozen in liquid nitrogen before being stored at -80°C prior to DNA
175 extraction. High molecular weight genomic DNA was isolated using a modified CTAB
176 procedure as previously described (Porebski et al. 1997). The quality of each extraction
177 was checked by pulsed-field electrophoresis (CHEF) on 1% agarose gels for size and
178 restriction enzyme digestibility, and quantified by Qubit fluorometry (Thermo Fisher
179 Scientific, Waltham, MA).

180 **Library construction and sequencing**

181 Genomic DNA from all 12 accessions were sequenced using the PacBio single-molecule
182 real-time (SMRT) platform, and the Illumina platform for genome size estimations and
183 sequence polishing. High molecular weight (HMW) DNA from each accession was
184 gently sheared into large fragments (*i.e.* 30Kb - 60Kb) using 26-gauge needles and then
185 end-repaired according to manufacturer's instructions (Pacific Biosciences). Briefly,
186 using a SMRTbell Express Template Prep Kit, blunt hairpins and sequencing adaptors
187 were ligated to HMW DNA fragments, and DNA sequencing polymerases were bound to
188 the SMRTbell templates. Size selection of large fragments (above 15Kb) was performed
189 using a BluPippin electrophoresis system (Sage Science). The libraries were quantified
190 using a Qubit Fluorometer (Invitrogen, USA) and the insert mode size was determined
191 using an Agilent fragment analyzer system with sizes ranging between 30Kb - 40Kb. The
192 libraries then were sequenced using SMRT Cell 1M chemistry version 3.0 on a PacBio
193 Sequel instrument. The number of long-reads generated per accession ranged from 2.01
194 million (LIMA::IRGC 81487-1) to 5.40 million (Azucena). The distribution of subreads
195 is shown in Figure S2 and the average lengths ranged from 10.58 Kb (Azucena) to 20.61
196 Kb (LIMA::IRGC 81487-1) (Table 2). According to the estimated genome size of the
197 IRGSP RefSeq, the average PacBio sequence coverage for each accession varied from
198 103x (LIMA::IRGC 81487-1) to 149x (IR 64) (Table 2).

199 For Illumina short-read sequencing, HMW DNA from each accession was sheared to
200 between 250-1000bp, followed by library construction targeting 350bp inserts following
201 standard Illumina protocols (San Diego, CA, USA). Each library was 2 x 150bp paired-

202 end sequenced using an Illumina X-ten platform. Low-quality bases and paired reads
203 with Illumina adaptor sequences were removed using *Trimmomatic* (Bolger et al. 2014).
204 Quality control for each library data set was carried out with *FastQC* (Brown et al. 2017).
205 Finally, between 36.52-Gb and 51.05-Gb of clean data from each accession was
206 generated and used for genome size estimation (Table S1) by Kmer analysis (Figure S3)
207 and the Genome Characteristics Estimation (GCE) program (Liu et al. 2013).

208 **Bionano optical genome maps**

209 Bionano optical maps for each accession were generated as previously described (Ou et
210 al. 2019), except that ultra-HMW DNA isolation, from approximately 4g of flash-frozen
211 dark-treated (48 hour) leaf tissue per accession, was performed according to a modified
212 version of the protocol described by Luo and Wing (Luo and Wing, 2003). Prior to
213 labeling, agarose plugs were digested with agarase and the starch and debris removed by
214 short rounds of centrifugation at 13,000 X g. DNA samples were further purified and
215 concentrated by drop dialysis against TE Buffer. Data processing, optical map assembly,
216 hybrid scaffold construction and visualization were performed using the Bionano Solve
217 (Version 3.4) and Bionano Access (v12.5.0) software packages
218 (<https://bionanogenomics.com/>).

219 ***De novo* genome assembly**

220 Genome assembly for each of the 12 genomes followed a five-step procedure as shown in
221 (Figure 2):

222 Step 1: PacBio subreads were assembled *de novo* into contigs using three genome
223 assembly programs: FALCON (Chin et al. 2016), MECAT2 (Xiao et al. 2017) and
224 Canu1.5 (Koren et al. 2017). The number of *de novo* assembled contigs obtained varied
225 from 51 (e.g. NATEL BORO::IRGC 34749-1 and KETAN NANGKA::IRGC 19961-2)
226 to 1,473 (CHAO MEO::IRGC 80273-1) for the 12 genomes (Table S2).

227 Step 2: Genome Puzzle Master (GPM) software (Zhang et al. 2016c) was used to merge
228 the *de novo* assembled contigs from the three assemblers, using the high-quality *O. sativa*
229 vg. indica cv. Minghui 63 reference genome MH63RS2 (Zhang et al. 2016a,b) as a guide.
230 GPM is a semi-automated pipeline that is used to integrate logical relationship data (*i.e.*
231 contigs from three assemblers for each accession) based on a reference guide. Contigs
232 were merged in the ‘assemblyRun’ step, with default parameters (minOverlapSeqToSeq
233 was set at 1 Kb and identitySeqToSeq was set at 99%). Redundant overlapping sequences

234 were also removed for each assembled contig. In addition, we gave contiguous contigs a
235 higher priority than ones with gaps to be retained in each assembly. After manual
236 checking, editing, and redundancy removal, the number of contigs in each assembly
237 ranged from 26 (NATEL BORO::IRGC 34749-1) to 588 (LIU XU::IRGC 109232-1)
238 (Table S3).

239 Step 3: The sequence quality of each contig was then improved by “sequence polishing”:
240 twice with PacBio long reads and once with Illumina short reads. Briefly, PacBio
241 subreads were aligned to GPM edited contigs using the software *blasr* (Chaisson and
242 Tesler 2012). All default parameters were used, except minimum align length, which was
243 set to 500-bp. Secondly, the tool *arrow* as implemented in SMRTlink6.0 (Pacific
244 Biosciences of California, Inc) was used for polishing the GPM edited contigs. The *bwa-*
245 *mem* program (Li 2013) was then used for mapping short Illumina reads onto assembled
246 contigs, and the tool *pilon* (Walker et al. 2014) was used for a final polishing step with
247 default settings.

248 Step 4: The polished contigs for each accession were arranged into pseudomolecules
249 using *GPM*, using MH63RS2 (Zhang et al. 2016a,b) as the reference guide. The program
250 *blastn* (Altschul et al. 1997) with a minimum alignment length of 1 Kb and an e-value <
251 $1e^{-5}$ as the threshold was used to align the corrected contigs to the reference guide. In
252 doing so, the corrected contigs were assigned to chromosomes, as well as ordered and
253 orientated in the GPM assembly viewer function. The number of contigs after step 4
254 ranged from a minimum of 15 contigs (GOBOL SAIL (BALAM)::IRGC 26624-2) to a
255 maximum of 104 contigs (IR64) (Table 3). The assembly size for the 12 accessions
256 ranged from 376.86 Mb (CHAO MEO::IRGC 80273-1) to 393.74 Mb (KHAO YAI
257 GUANG::IRGC 65972-1) (Table 3) and the length of individual chromosome varied
258 from 23.06 Mb (chromosome 9 of CHAO MEO::IRGC 80273-1) to 44.96 Mb
259 (chromosome 1 of LIMA::IRGC 81487-1) (Table S4). The average N50 value was 23.10
260 Mb, with the highest and the lowest values being 30.91 Mb (LIU XU::IRGC 109232-1)
261 and 7.35 Mb (IR 64), respectively. The average number of gaps among the 12 new
262 genome assemblies was 18, with 8 assemblies containing less than 10 gaps (Table 3).

263 Step 5: To independently validate our assemblies, we generated and compared Bionano
264 optical maps to each assembly. In total, 17 (Azucena) to 56 (LIU XU::IRGC 109232-1)
265 Bionano optical maps were constructed for all 12 rice accessions, which yielded contig
266 N50 values of between 22.75 Mb (CHAO MEO::IRGC 80273-1) to 31.45 Mb (KHAO
267 YAI GUANG::IRGC 65972-1) (Table S5). As shown in Figure 3 and Figure S4, the

268 chromosomes and/or chromosome arms of all 12 *de novo* assemblies were highly
269 supported by these ultra-long optical maps. Although rare, a few discrepancies between
270 the optical maps and genome assemblies can be seen and are likely due to small errors
271 and chimeras that can be produced through both the optical mapping and sequence
272 assembly pipelines (Udall and Dawe 2018).

273 Following these five steps, we were able to produce 12 near-gap-free *Oryza sativa*
274 platinum standard reference genome sequences (PSRefSeqs) that represent 12 of 15
275 subpopulations of cultivated Asian rice.

276 **BUSCO evaluation**

277 The Benchmarking Universal Single-Copy Orthologs (BUSCO3.0) software package
278 (Simao et al. 2015) was employed to evaluate the gene space completeness of the 12
279 genome assemblies. These genomes captured, on average, 97.9% of the BUSCO
280 reference gene set, with a minimum of 95.7% (IR64) and a maximum of 98.6% (LARHA
281 MUGAD::IRGC 52339-1 and KHAO YAI GUANG::IRGC 65972-1) (Table 3).

282 Of note, when performing this analysis, we observed that on average 30 out of the
283 1,440 conserved BUSCO genes tested (<https://www.orthodb.org/v9/index.html>) were
284 missing from each new assembly, 16 of which were not present in all 12, plus the IRGSP,
285 ZS 97, MH 63 and N 22 RefSeqs (Figure S5). This result suggested that these 16
286 “conserved” genes may not exist in rice, or other cereal genomes, thereby artificially
287 reducing the BUSCO gene space scores for our 12 assemblies. To test this hypothesis, we
288 searched for all 16 genes missing in maize, which diverged from rice about 50 million
289 years ago (MYA) (Wolfe et al., 1989, Gale et al., 1998 and Guo et al., 2019). We found
290 that 13 of the 16 genes in question could not be found in 3 high-quality recently
291 published maize genome assemblies (Figure S5) and therefore, concluded that 13 of the
292 16 “conserved” genes in the BUSCO database are not present in cereals, and should be
293 excluded from our gene space analysis. Taking this into account, we recalculated the
294 BUSCO gene space content for each of 12 assemblies and found that 10 of 12 assemblies
295 captured more than 98% of the BUSCO gene set (Table 3).

296 **Transposable element (TE) prediction**

297 To determine the pan-transposable element content of cultivated Asian rice we analyzed
298 the 12 new reference genomes, presented here, along with the MH 63, ZS 97, N 22
299 PacBio reference genomes. In addition, we also included a reanalysis of the IRGSP

300 RefSeq as it is conventionally considered the standard rice genome for which all
301 comparisons are conducted. This 16 genome data set was used to represent the K=15
302 population structure of cultivated Asian rice.

303 A search for sequences similar to TEs was carried out using RepeatMasker (Smit
304 AFA et al, 2013) run under default parameters with the exception of the options: -
305 no_is -nolow. RepeatMasker was run using the library “rice 7.0.0.liban”, which is
306 an updated in-house version of the publicly available MSU_6.9.5 library (Ou et al. 2019),
307 retrieved from
308 https://github.com/oushujun/EDTA/blob/master/database/Rice_MSU7.fasta.std6.9.5.out.
309 The average TE content of this 16 genome data set was 47.66% with a minimum value of
310 46.07% in IRGSP RefSeq and a maximum of 48.27% in KHAO YAI GUANG::IRGC
311 65972-1 (Table 4). The major contribution to this fraction was composed of long terminal
312 repeat retrotransposons (LTR-RTs, min: 23.55%, max: 27.27% and average: 25.96%)
313 followed by DNA-TEs (min:14.87%, max, 16.18% and average: 15.26%). Long
314 interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)
315 were identified as on average 1.43% and 0.39% of the 16 genomes, respectively.

316 **Structural Variants**

317 Each genome assembly (n=16), as described above, was fragmented using the EMBOSS
318 tool *splitter* (Rice et al. 2000) to create a 10x genome equivalent redundant set of 50kb
319 reads. These reads were then mapped onto every other genome assembly using the tool
320 *NGMLR* (Sedlazeck et al. 2018). Finally, the software *SVIM* (Heller and Vingron 2019)
321 was run under default parameters to parse the mapping output. Only insertions, deletions
322 and tandem duplications up to a maximum length of 25 Kb were considered in this
323 analysis.

324 The results of this analysis identified several thousand insertions and deletions
325 whenever an assembly was compared to any other. Greater variability was found between
326 varieties belonging to different major groups (e.g. *Geng-japonica* vs. *Xian-indica*) than
327 occurred between those within these groups. The amount of genome sequences with
328 structural variation between any two varieties ranged from 17.57 Mb to 41.54 Mb for
329 those belonging to the indica (XI, *Xian-indica*) varietal group (avg: 31.75 Mb) and from
330 18.55 Mb to 23.07 Mb (avg: 21.00 Mb) for those in the japonica (GJ, *Geng-japonica*)
331 varietal group. When all 16 genomes are considered together, the range is between 17.57
332 Mb and 41.54 Mb, with an average value of 33.70 Mb (Table S6). The total unshared

333 fraction collected out of all pairwise comparisons was composed for 89.89% by TE
334 related sequences.

335 **Data Records**

336 Data for all 12 genome shotgun sequencing projects have been deposited in Genbank
337 (<https://www.ncbi.nlm.nih.gov/>), including PacBio raw data, Illumina raw data, Bionano
338 optical maps and the twelve PSRefSeqs. The BioProjects, BioSamples, Genome
339 assemblies, Sequence Read Archives (SRA) accession and supplementary files (Bionano
340 optical maps) of 12 genomes are listed in Table 3.

341 **Technical Validation**

342 DNA sample quality

343 DNA quality was checked by pulsed-field gel electrophoresis for size and restriction
344 enzyme digestibility. Nucleic acid concentrations were quantified by Qubit fluorometry
345 (Thermo Fisher Scientific, Waltham, MA).

346

347 Illumina libraries

348 Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit
349 for Illumina Libraries (KapaBiosystems, Wilmington, MA, USA), and library profiles
350 were evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara,
351 CA, USA).

352

353 Gene Space Completeness

354 Benchmarking Universal Single-Copy Orthologs (BUSCO3.0) was executed using the
355 `embryophyta_odb9.tar.gz` database to assess the gene space of each genome, minus 13
356 genes that do not appear to exist in the cereal genomes tested (Figure S5).

357

358 Assembly accuracy

359 Bionano optical maps were generated and used to validate all 12 genome assemblies.

360

361 This paper is the first release of 12 PSRefSeqs, optical maps and all associated raw data
362 for the accessions listed in Table 3.

363 **Code Availability**

364 The population re-analysis of 3K-RG dataset and 12 genome assemblies were obtained
365 using several publicly available software packages. To allow researchers to precisely
366 repeat any steps, the settings and the parameters used are provided below:

367

368 Population structure

369 ADMIXTURE (Alexander et al. 2009) was run with default options. The R scripts for
370 further population structure analysis, including setting up CLUMPP files, can be found in
371 Github repository <https://github.com/dchebotarov/Q-aggr>.

372

373 Genome size estimation:

374 The K-mer and GCE program (Liu et al. 2013) were employed for genome size
375 estimation. Command line:

```
376 kmer_freq_hash -k (13-17) -l genome.list -a 10 -d 10 -t 8 -  
377 i 400000000 -o 0 -p genom_kmer(13-17) &> genome_kmer(13-  
378 17)_freq.log, and gce -f genom_kmer(13-17).freq.stat -c  
379 $peak -g #amount -m 1 -D 8 -b 1 -H 1 > genome.table 2>  
380 genom_kmer(13-17).log
```

381

382 Genome assembly:

383 (1) *MECAT2*: all parameters were set to the defaults. Command line:

```
384 mecat.pl config_file.txt, mecat.pl correct config_file.txt  
385 and mecat.pl assemble config_file.txt
```

386 (2) *Canu1.5*: all parameters were set to the defaults. Command line:

```
387 canu -d canu -p canu genomeSize=400m -pacbio-raw  
388 rawreads.fasta
```

389 (3) *FALCON*: all parameters were set to the defaults. Command line:

```
390 fc_run.py fc_run.cfg &>fc_run.out
```

391 (4) *GPM*: manual edit with merging *de novo* assemblies from *MECAT2*, *Canu1.5*, and
392 *FALCON*.

393 Polishing:

394 (1) *arrow*: all parameters were set to the defaults except alignment length = 500
395 bp. The *arrow* polish was carried out by the SMRT Link v6.0 webpage
396 (<https://www.pacb.com/support/software-downloads/>).

397 (2) *pilon1.18*: all parameters were set to the defaults and *pilon* polish was carried out as
398 recommended at the SMRT Link v6.0 ([https://www.pacb.com/support/software-](https://www.pacb.com/support/software-downloads/)
399 [downloads/](https://www.pacb.com/support/software-downloads/)).

400

401 BUSCO:

402 The BUSCO3.0 version was employed in this study. Command line: `run_BUSCO.py`
403 `-i genome.fasta -o genome -l embryophyta_odb9 -m genome -c`
404 `16`

405

406 RepeatMasker:

407 The repeat sequences were employed with the library `rice7.0.0_liban` in-house. Command
408 line: `RepeatMasker -pa 24 -x -no_is -nolow -cutoff 250 -lib`
409 `rice7.0.0_liban.txt genome.fasta`

410 **Acknowledgements**

411 This research was supported by the AXA Research Fund (International Rice Research
412 Institute), the King Abdullah University of Science & Technology, and the Bud Antle
413 Endowed Chair for Excellent in Agriculture (University of Arizona) to R.A.W., the Start-
414 up Fund of Huazhong Agricultural University to J.Z., and funding from the Taiwan
415 Council of Agriculture to IRRI. The BUSCO analysis data for maize was kindly
416 provided by Dr. Wu and Dr. Li from the Institute of Plant Physiology and
417 Ecology, and Dr. Wang from Shanghai Jiao Tong University. One of two TE
418 libraries used for repeat analysis was provided by Dr. Eric Laserre (University of
419 Perpignan, France)

420 **Author contributions**

421 J.Z., K.M., D.C., M.L., N.A., N.R.S.H., H.L., R.M, and R.A.W. designed and conceived
422 the research. D.C. and K.M. perform the population structure analysis. K.M., M.L.,
423 L.J.A., N.L. generated and provided SSD seed 12 *O. sativa* accessions. D.K., S.L., S.R.,

424 N.M prepared DNA and performed PacBio and Illumina sequencing. C.S.-S. managed all
425 PacBio and Illumina sequence data processing and transfer. P.P. and V.L. generated all
426 Bionano optical maps. J.Z. and Y.Z. performed sequence assembly. Y.Z. carried out
427 genome size estimation, GPM editing, assembly polishing and data submission. V.L. and
428 Y.Z. analyzed the Bionano optical maps and the validation of 12 PSRefSeqs. A.Z. and
429 Y.Z. carried out TE prediction and structural analysis. Y.Z., N.A., A.Z., J.Z., D.C., M.L.,
430 K.M., N.M. and R.A.W. wrote and edited the paper. All authors read and approved the
431 final manuscript.

432 **Competing interests**

433 The authors declare that there is no conflict of interest regarding the publication of this
434 article.

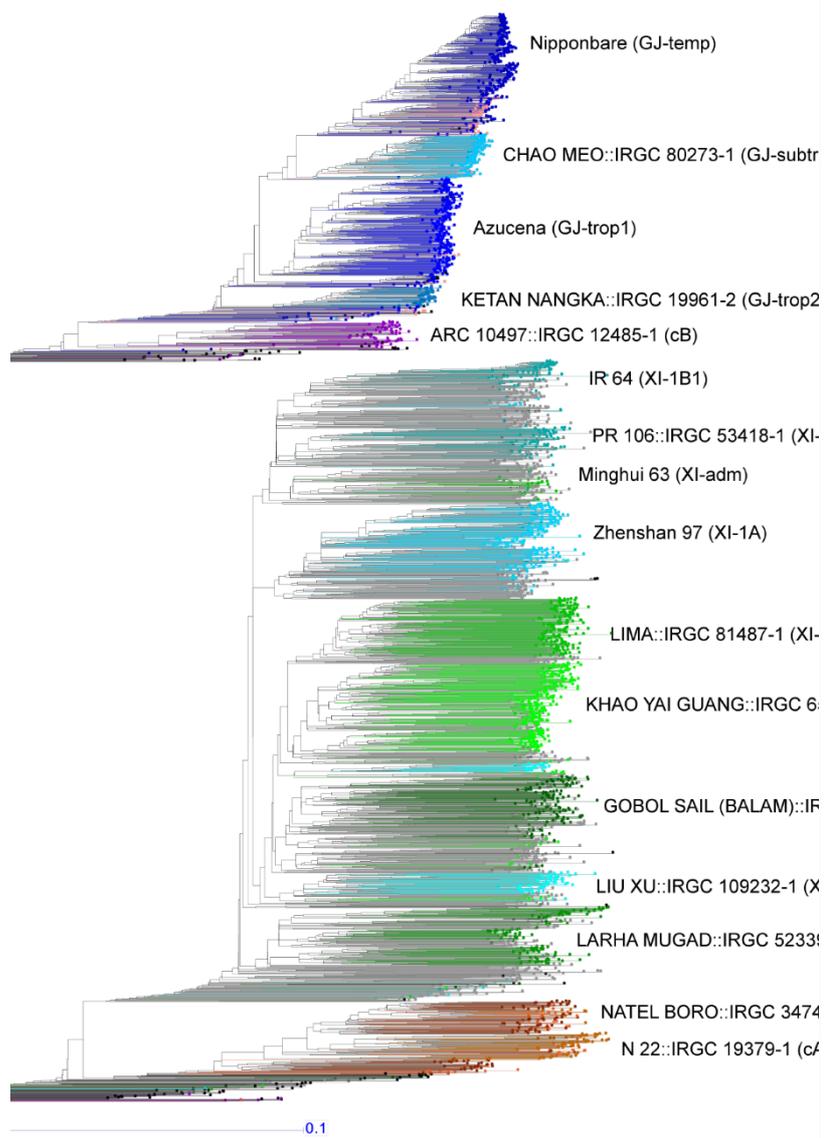
435 **Figure legends**

436 **Figure 1.** Phylogenetic tree with the accession selected for PSRefSeq sequencing for
437 each of the K=15 subpopulations and a single admixture group. Groups are colored
438 according to the assignment from Admixture analysis. The subpopulation designation is
439 in parentheses following the name.

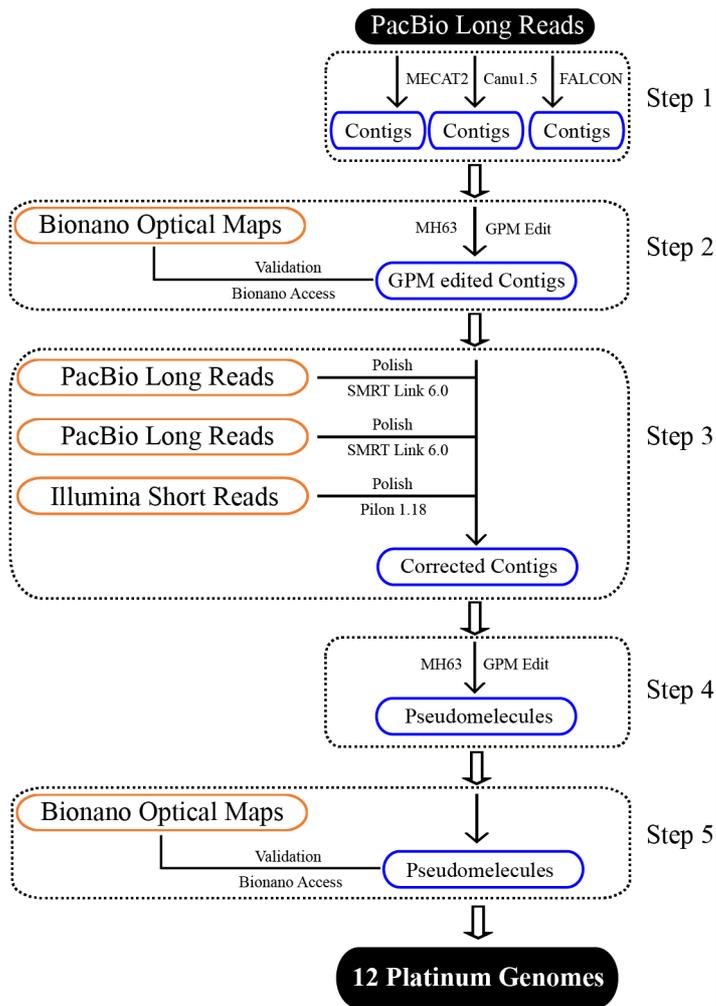
440 **Figure 2.** Genome assembly and validation pipeline.

441 **Figure 3.** Bionano optical map validation of chromosome 1 for 12 *de novo* assemblies.

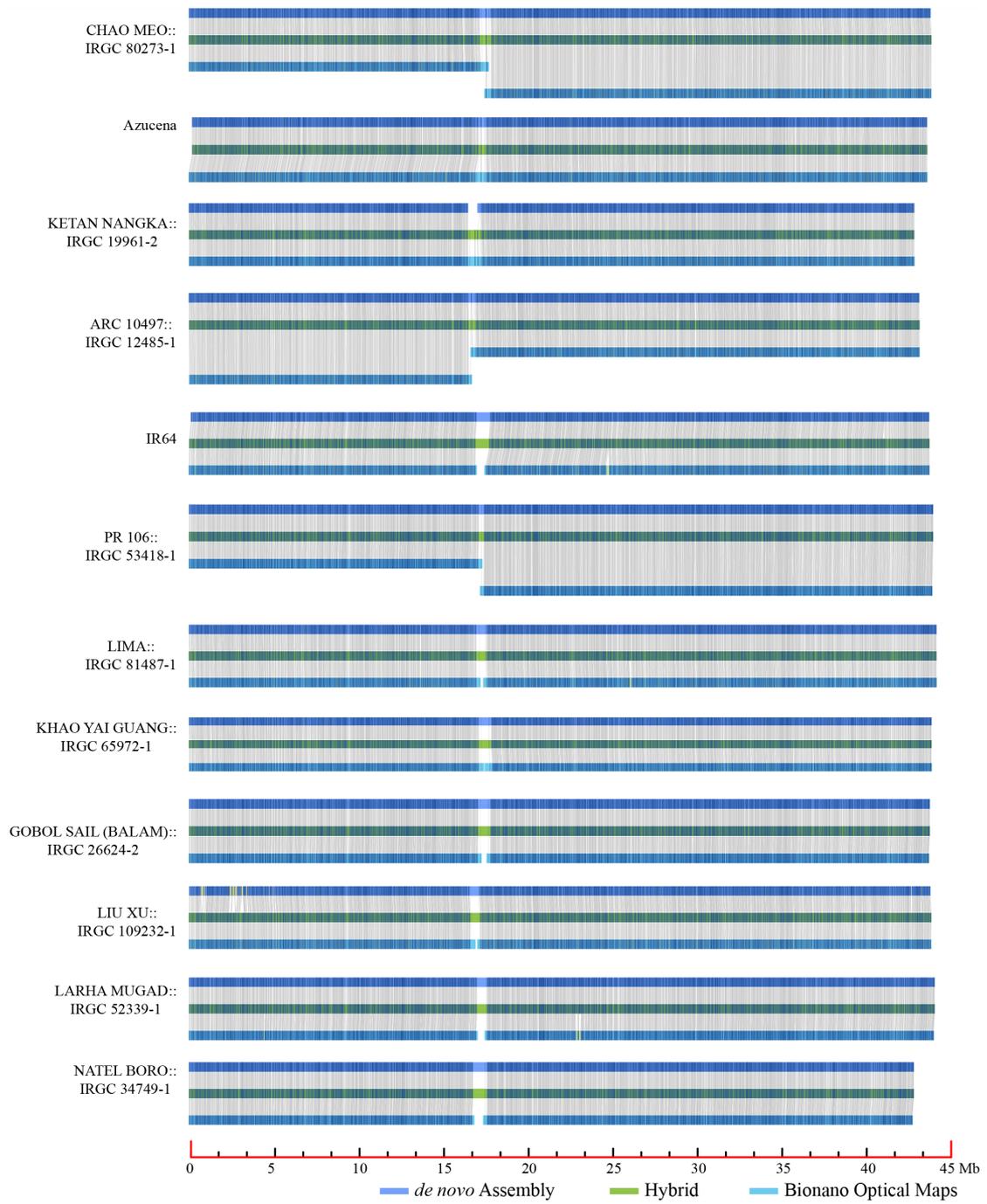
442 Figure 1



443 Figure 2



444 Figure 3



445 **Tables**

446 Table 1. Sample collection information for 12 *Oryza sativa* accessions.

447 Table 2. Sequencing platforms used and data statistics for the 12 *Oryza sativa* genomes.

448 Table 3. *de novo* assembly, BUSCO evaluation and accession numbers in GenBank of the

449 12 *Oryza sativa* genomes.

450 Table 4. Abundance of the major TE classes in the 16 *Oryza sativa* genomes.

451 Table 1.

Variety Name	Genetic Stock ID	Country Origin	12 subpopulations
CHAO MEO::IRGC 80273-1	IRGC 132278	Lao PDR	GJ-subtrp
Azucena	I1A41685	Philippines	GJ-trop1
KETAN NANGKA::IRGC 19961-2	IRGC 128077	Indonesia	GJ-trop2
ARC 10497::IRGC 12485-1	IRGC 117425	India	cB
IR 64	I1A42114	Philippines	XI-1B1
PR 106::IRGC 53418-1	IRGC 127742	India	XI-1B2
LIMA::IRGC 81487-1	IRGC 127564	Indonesia	XI-3A
KHAO YAI GUANG::IRGC 65972-1	IRGC 127518	Thailand	XI-3B1
GOBOL SAIL (BALAM)::IRGC 26624-2	IRGC 132424	Bangladesh	XI-2A
LIU XU::IRGC 109232-1	IRGC 125827	China	XI-3B2
LARHA MUGAD::IRGC 52339-1	IRGC 125619	India	XI-2B
NATEL BORO::IRGC 34749-1	IRGC 127652	Bangladesh	cA2

Subpopulations: GJ = *Geng-japonica* where trop = tropical, subtrp = subtropical; cB = *circum-Basmati*; XI = *Xian-indica*; cA = *circum-Aus*

452 Table 2.

Variety Name	Sequencing platform	Raw data (Gb)	Depth	Number of subreads (M)	Mean subread length (Kb)
CHAO MEO::IRGC 80273-1	PacBio Sequel	49.1	123X	4.26	11.526
Azucena	PacBio Sequel	57.1	143X	5.40	10.581
KETAN NANGKA::IRGC 19961-2	PacBio Sequel	49.8	125X	2.78	17.876
ARC 10497::IRGC 12485-1	PacBio Sequel	44.7	112X	4.06	11.026
IR 64	PacBio Sequel	59.7	149X	5.24	11.393
PR 106::IRGC 53418-1	PacBio Sequel	42.2	105X	2.08	20.317
LIMA::IRGC 81487-1	PacBio Sequel	41.4	103X	2.01	20.612
KHAO YAI GUANG::IRGC 65972-1	PacBio Sequel	42.5	106X	2.37	17.954
GOBOL SAIL (BALAM)::IRGC 26624-2	PacBio Sequel	42.2	105X	2.13	19.777
LIU XU::IRGC 109232-1	PacBio Sequel	55.3	138X	3.66	15.109
LARHA MUGAD::IRGC 52339-1	PacBio Sequel	45.1	113X	3.22	14.011
NATEL BORO::IRGC 34749-1	PacBio Sequel	44.4	111X	2.74	16.2

453 Table 3.

Variety Name	BioProject	BioSample	Genome size (bp)	# Contigs	Contig N50 (Mb)	# Gaps	Scaffold N50 (Mb)	BUSCO	Adjust BUSCO	Genome Accession	SRP	Supplementary Files (Bionano optical map)
CHAO MEO::IRGC 80273-1	PRJNA565484	SAMN12748601	376,856,903	55	11.02	43	30.35	97.60%	98.49%	VYIH00000000	SRP226088	SUPPF_0000003210
Azucena	PRJNA424001	SAMN08217222	379,627,553	28	22.94	16	30.95	97.80%	98.69%	PKQC00000000	SRP227255	SUPPF_0000003212
KETAN NANGKA::IRGC 19961-2	PRJNA564615	SAMN12718029	380,759,091	21	22.68	9	30.70	98.00%	98.89%	VYIC00000000	SRP226080	SUPPF_0000003204
ARC 10497::IRGC 12485-1	PRJNA565479	SAMN12748569	378,463,869	40	17.92	28	30.57	98.40%	99.30%	VYID00000000	SRP226093	SUPPF_0000003206
IR 64	PRJNA509165	SAMN10564385	386,698,898	104	7.35	92	31.22	95.70%	96.57%	RWKJ00000000	SRP227298	SUPPF_0000003213
PR 106::IRGC 53418-1	PRJNA563359	SAMN12672924	391,176,105	16	27.05	4	32.03	96.60%	97.48%	VYIB00000000	SRP226078	SUPPF_0000003202
LIMA::IRGC 81487-1	PRJNA564572	SAMN12715984	392,625,308	17	27.37	5	32.42	98.50%	99.40%	VXJH00000000	SRP226079	SUPPF_0000003203
KHAO YAI GUANG::IRGC 65972-1	PRJNA565481	SAMN12748590	393,737,720	19	21.82	7	32.08	98.60%	99.50%	VYIF00000000	SRP226086	SUPPF_0000003208
GOBOL SAIL (BALAM)::IRGC 26624-2	PRJNA564763	SAMN12721963	391,772,995	15	29.60	3	31.75	97.90%	98.79%	VXJI00000000	SRP226082	SUPPF_0000003205
LIU XU::IRGC 109232-1	PRJNA577228	SAMN13021815	392,033,263	17	30.91	5	32.30	98.40%	99.30%	WGGU00000000	SRP226085	SUPPF_0000003211
LARHA MUGAD::IRGC 52339-1	PRJNA565480	SAMN12748589	390,195,943	16	30.75	4	32.10	98.60%	99.50%	VYIE00000000	SRP226084	SUPPF_0000003207
NATEL BORO::IRGC 34749-1	PRJNA565483	SAMN12748600	383,720,936	16	27.83	4	31.31	98.10%	98.99%	VYIG00000000	SRP226087	SUPPF_0000003209

454 Table 4.

Variety Name	TOTAL	LTR-RT	LINEs	SINEs	DNA_TEs	Unclassified
NIPPONBARE	46.07	23.55	1.52	0.41	16.18	4.41
CHAO MEO::IRGC 80273-1	46.25	24.00	1.46	0.40	15.59	4.80
Azucena	47.07	24.48	1.47	0.40	15.82	4.89
KETAN NANGKA::IRGC 19961-2	46.99	24.87	1.47	0.40	15.72	4.53
ARC 10497::IRGC 12485-1	46.95	24.74	1.48	0.40	15.68	4.65
IR 64	47.87	26.82	1.42	0.40	14.97	4.26
PR 106::IRGC 53418-1	47.95	26.82	1.41	0.39	15.05	4.28
Minghui 63	47.97	26.61	1.44	0.4	15.3	4.22
Zhenshan 97	47.95	26.79	1.42	0.39	15.19	4.16
LIMA::IRGC 81487-1	48.04	26.87	1.40	0.39	15.01	4.37
KHAO YAI GUANG::IRGC 65972-1	48.27	27.27	1.40	0.39	14.87	4.34
GOBOL SAIL (BALAM)::IRGC 26624-2	48.15	26.99	1.40	0.39	14.99	4.38
LIU XU::IRGC 109232-1	46.92	27.06	1.26	0.32	14.31	3.97
LARHA MUGAD::IRGC 52339-1	48.05	26.74	1.41	0.39	15.09	4.42
NATEL BORO::IRGC 34749-1	47.33	25.75	1.42	0.40	15.12	4.64
N 22::IRGC 19379-1	47.79	25.95	1.44	0.39	15.20	4.81

455 **References**

- 456 3K RGP. The 3,000 rice genomes project. *GigaScience* 3.1 (2014): 2047-217X.
- 457 Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search
458 programs. *Nucleic acids research* 25.17 (1997): 3389-3402.
- 459 Bolger, A. M., Marc L., and Bjoern U. Trimmomatic: a flexible trimmer for Illumina sequence
460 data. *Bioinformatics* 30.15 (2014): 2114-2120.
- 461 Brown, J., Meg P., and Lee A.M. FQC Dashboard: integrates FastQC results into a web-based,
462 interactive, and extensible FASTQ quality control tool. *Bioinformatics* 33.19 (2017): 3137-3139.
- 463 Chaisson, M.J., and Glenn T. Mapping single molecule sequencing reads using basic local alignment
464 with successive refinement (BLASR): application and theory. *BMC bioinformatics* 13.1 (2012):
465 238.
- 466 Chin, C. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature*
467 *methods* 13.12 (2016): 1050.
- 468 Alexander, D.H., John N., and Kenneth L. Fast model-based estimation of ancestry in unrelated
469 individuals. *Genome research* 19.9 (2009): 1655-1664.
- 470 Gale, M.D., and Katrien M.D. Comparative genetics in the grasses. *Proceedings of the National*
471 *Academy of Sciences* 95.5 (1998): 1971-1974.
- 472 Guo, H. *et al.* Gene duplication and genetic innovation in cereal genomes. *Genome research* 29.2
473 (2019): 261-269.
- 474 Heller, D., and Martin V. SVIM: structural variant identification using mapped long
475 reads. *Bioinformatics* 35.17 (2019): 2907-2915.
- 476 Huang, X.H. *et al.* A map of rice genome variation reveals the origin of cultivated
477 rice. *Nature* 490.7421 (2012): 497.
- 478 International Rice Genome Sequencing Project. The map-based sequence of the rice
479 genome. *Nature* 436.7052 (2005): 793.
- 480 Jakobsson, M., and Noah A.R.. CLUMPP: a cluster matching and permutation program for dealing
481 with label switching and multimodality in analysis of population structure. *Bioinformatics* 23.14
482 (2007): 1801-1806.
- 483 Kawahara, Y. *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next
484 generation sequence and optical map data. *Rice* 6.1 (2013): 4.
- 485 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and
486 repeat separation. *Genome research* 27.5 (2017): 722-736.
- 487 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
488 *preprint arXiv:1303.3997* (2013).
- 489 Li, J.Y., Wang J., and Robert S.Z. The 3,000 rice genomes project: new opportunities and challenges
490 for future rice research. *GigaScience* 3.1 (2014): 8.
- 491 Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome
492 projects. *arXiv preprint arXiv:1308.2012* (2013).
- 493 Luo, M., and Wing. A.R. An improved method for plant BAC library construction. *Plant functional*
494 *genomics*. Humana Press, 2003. 3-19.
- 495 McNally, K. L. *et al.* Genomewide SNP variation reveals relationships among landraces and modern
496 varieties of rice. *Proceedings of the National Academy of Sciences* 106.30 (2009): 12273-12278.
- 497 Merrey, D. J. *et al.* Agricultural Development and Sustainable Intensification. Routledge, 2018.

498 Ou, S. *et al.* Effect of sequence depth and length in long-read assembly of the maize inbred
499 nc358. *BioRxiv* (2019): 858365.

500 Ou, S. *et al.* Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined,
501 Comprehensive Pipeline. *bioRxiv* (2019): 657890.

502 Porebski, S., Bailey, L. G., & Baum, B. R. Modification of a CTAB DNA extraction protocol for
503 plants containing high polysaccharide and polyphenol components. *Plant molecular biology*
504 *reporter* 15.1 (1997): 8-15.

505 Rhoads, A., and Kin F.A. PacBio sequencing and its applications. *Genomics, proteomics &*
506 *bioinformatics* 13.5 (2015): 278-289.

507 Rice, P., Ian L., and Alan B. EMBOSS: the European molecular biology open software suite. (2000):
508 276-277.

509 Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule
510 sequencing. *Nature methods* 15.6 (2018): 461.

511 Simão, F. A., *et al.* BUSCO: assessing genome assembly and annotation completeness with single-
512 copy orthologs. *Bioinformatics* 31.19 (2015): 3210-3212.

513 Stein, J. C., *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation,
514 turnover and innovation across the genus *Oryza*. *Nature genetics* 50.2 (2018): 285.

515 Maja T. and Chen N. Using RepeatMasker to identify repetitive elements in genomic
516 sequences. *Current protocols in bioinformatics* 25.1 (2009): 4-10.

517 Thomson, M J. *et al.* Large-scale deployment of a rice 6 K SNP array for genetics and breeding
518 applications. *Rice* 10.1 (2017): 40.

519 Udall, J. A., and Kelly D. Is it ordered correctly? Validating genome assemblies by optical
520 mapping. *The Plant Cell* 30.1 (2018): 7-14.

521 Walker, B. J., *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and
522 genome assembly improvement. *PloS one* 9.11 (2014): e112963.

523 Wang, W. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated
524 rice. *Nature* 557.7703 (2018): 43.

525 Wing, A.W., Michael D. P., and Zhang Q.F. The rice genome revolution: from an ancient grain to
526 Green Super Rice. *Nature Reviews Genetics* 19.8 (2018): 505-517.

527 Wolfe, K. H. *et al.* Date of the monocot-dicot divergence estimated from chloroplast DNA sequence
528 data. *Proceedings of the National Academy of Sciences* 86.16 (1989): 6201-6205.

529 Xiao, C. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule
530 sequencing reads. *nature methods* 14.11 (2017): 1072.

531 Zhang, J. *et al.* Building two indica rice reference genomes with PacBio long-read and Illumina
532 paired-end sequencing data. *Scientific data* 3 (2016): 160076.

533 Zhang, J. *et al.* Extensive sequence divergence between the reference genomes of two elite indica rice
534 varieties Zhenshan 97 and Minghui 63. *Proceedings of the National Academy of Sciences* 113.35
535 (2016): E5163-E5171.

536 Zhang, J. *et al.* Genome puzzle master (GPM): an integrated pipeline for building and editing
537 pseudomolecules from fragmented sequences. *Bioinformatics* 32.20 (2016): 3058-3064.

538 Zhao, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild
539 rice. *Nature genetics* 50.2 (2018): 278.

540 **Supplementary Information**

541 **Supplementary file1**

542 **Supplementary Table 1.** Summary of Illumina genome survey sequences for 12 *Oryza*
543 *sativa* genomes.

544 **Supplementary Table 2.** Genome features of *de novo* assemblies for 12 *Oryza sativa*
545 accessions by Canu1.5, FALCON and MECAT2.

546 **Supplementary Table 3.** Genome features of 12 *Oryza sativa* accessions by GPM
547 editing.

548 **Supplementary Table 4.** Chromosome length (Mb) of 12 *Oryza sativa* genomes.

549 **Supplementary Table 5.** Bionano optical map statistics of 12 *Oryza sativa* genomes.

550 **Supplementary Table 6.** Summary of large structural variation (>50 bp) by comparison
551 of each of 16 genomes to every other genome (including 12 genomes from this study and
552 4 previously reported: MH63, ZS97, N 22 and the IRGSP RefSeq).

553 **Supplementary file2**

554 **Supplementary Figure 1.** Admixture results for K=5 to 15. The samples are grouped
555 according to the new classification. At K=9,12,13, the Q matrices converged to two
556 different modes, differing according to whether ind1A is split, or tropical japonica.

557 **Supplementary Figure 2.** Length distribution of PacBio long reads used for 12 *Oryza*
558 *sativa* genome assemblies.

559 **Supplementary Figure 3.** K-mer analysis of Illumina short sequences that were used for
560 genome size estimation with the GCE program.

561 **Supplementary Figure 4.** Bionano Access visualization view for 12 *de novo* assemblies
562 with Bionano optical maps and their underlying alignments.

563 **Supplementary Figure 5.** Summary of missing genes in the BUSCO gene space
564 evaluation of 12 *de novo Oryza sativa* assemblies, 4 public *Oryza sativa* PSRefSeqs and
565 3 high-quality *Zea mays* genomes.