1   Genome sequence of the model rice variety KitaakeX

2   Rashmi Jain[1,2], Jerry Jenkins[3,4], Shengqiang Shu[3], Mawsheng Chern[1,2], Joel A. Martin[3], Dario Copetti[7,8,9],

3   Phat Q. Duong[1,2], Nikki T. Pham[1], David A. Kudrna[7,8], Jayson Talag[7,8], Wendy S. Schackwitz[3], Anna M.

4   Lipzen[3], David Dilworth[3], Diane Bauer[3], Jane Grimwood[3,4], Catherine R. Nelson[1], Feng Xing[6], Weibo Xie[6],

5   Kerrie W. Barry[3], Rod A. Wing[7,8,9], Jeremy Schmutz[3,4], Guotian Li[1,2,5*], Pamela C. Ronald[1,2*]

6   [1]Department of Plant Pathology and the Genome Center, University of California, Davis, CA 95616, USA

7   [2]Feedstocks Division, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA

8   94720, USA

9   [3]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

10   [4]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA

11   [5]State Key Laboratory of Agricultural Microbiology and College of Plant Science and Technology,

12   Huazhong Agricultural University, Wuhan 430070, Hubei, China

13   [6]National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research

14   (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

15   [7]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

16   [8]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

17   [9]International Rice Research Institute, Genetic Resource Center, Los Baños, Laguna, Philippines

18   * co-corresponding authors

19   Address correspondence to pcronald@ucdavis.edu and li4@mail.hzau.edu.cn

20   **Corresponding authors:**   Prof. Pamela C. Ronald

21   Department of Plant Pathology, and the Genome Center, University of

22   California, Davis

23   One Shields Avenue, Davis, CA 95616, USA

24   E-mail: pcronald@ucdavis.edu

25   Tel: 1-530-752-1654

26   Fax: 1-530-752-6088

27   Prof. Guotian Li

28   State Key Laboratory of Agricultural Microbiolog and College of Plant

29   Science and Technologyy, Huazhong Agricultural University, Wuhan

30   430070, Hubei, China

31   E-mail: li4@mail.hzau.edu.cn

32

33

34 Abstract

35   Here, we report the *de novo* genome sequencing and analysis of *Oryza sativa* ssp. *japonica* variety

36   KitaakeX, a Kitaake plant carrying the rice XA21 immune receptor. Our KitaakeX sequence assembly

37   contains 377.6 Mb, consisting of 33 scaffolds (476 contigs) with a contig N50 of 1.4 Mb. Complementing

38   the assembly are detailed gene annotations of 35,594 protein coding genes. We identified 331,335 genomic

39   variations between KitaakeX and Nipponbare (ssp. *japonica*), and 2,785,991 variations between KitaakeX

40   and Zhenshan97 (ssp. *indica*). We also compared Kitaake resequencing reads to the KitaakeX assembly

41   and identified 219 small variations. The high-quality genome of the model rice plant KitaakeX will accelerate

42   rice functional genomics.

43   Keywords: Rice, Kitaake, KitaakeX, XA21 immune receptor, Whole genome sequence, *De novo* genome

44   assembly, Nipponbare, Zhenshan97

45 Background

46   Rice (*Oryza sativa*) provides food for more than half of the world's population [1] and also serves

47   as a model for studies of other monocotyledonous species. Cultivated rice contains two major types of *O.*

48   *sativa*, the *O. sativa indica/Xian* group and the *O. sativa japonica/Geng* group. Using genomic markers, two

49   additional minor types have been recognized, the circum-Aus group and the circum-Basmati group [2].

50   The Kitaake cultivar (ssp. *japonica*), which originated at the northern limit of rice cultivation in

51   Hokkaido, Japan [3], has emerged as a model for rice research [4] because it is extremely early flowering,

52   easy to propagate, and short in stature [5]. Kitaake has been used to establish multiple mutant

53   populations, including an RNAi mutant collection [6], T-DNA insertion collections [4], [7], and a whole-

54   genome sequenced mutant population of KitaakeX, a Kitaake variety carrying the *Xa21* immune receptor

55   gene (formerly called X.Kitaake) [8, 9]. Kitaake has been used to explore diverse aspects of rice biology,

56   including flowering time [10], disease resistance [11], [12], [13], small RNA biology [14], and the CRISPR-

57   Cas9 and TALEN technologies [15], [16].

58   The unavailability of the Kitaake genome sequence has posed an obstacle to the use of Kitaake

59   in rice research. For example, analysis of a fast-neutron (FN) induced mutant population in KitaakeX [8],

60   required the use of Nipponbare (ssp. *japonica)* as the reference. Additionally, CRISPR/Cas9 guide RNAs

61   cannot be accurately designed for Kitaake without a complete sequence. To address these issues, we

62   assembled a high-quality genome sequence of KitaakeX, compared its genome to the genomes of rice

63   varieties Nipponbare and Zhenshan97 (ssp. *indica*), and identified genomic variations.

Results

65   Kitaake has long been recognized as a rapid life-cycle variety [17], but it has yet to be systematically

66   compared to other rice varieties. We compared the flowering time of KitaakeX with other sequenced rice

67   varieties under long-day conditions (14 h light/10 h dark). Consistent with other studies, we found that

68   KitaakeX flowers much earlier than other varieties (Fig. 1a, 1b), heading at 54 days after germination. Other

69   rice varieties Nipponbare, 93-11 (ssp. *indica*), IR64 (ssp. *indica*), Zhenshan 97, Minghui 63 (ssp. *indica*),

70   and Kasalath (aus rice cultivar) start heading at 134, 99, 107, 79, 125, and 84 days after germination,

71   respectively (Fig. 1b).

72   We assessed how KitaakeX is related to other rice varieties using a phylogenetic approach based

73   on the rice population structure and diversity published for 3,010 varieties [2]. The 3010 sequenced

74   accessions were classified into nine subpopulations, most of which could be connected to geographical

75   origins. The phylogenetic tree reveals that KitaakeX and Nipponbare are within the same subpopulation

76   closely related (Fig 1c).

3

77   To obtain a high-quality, *de novo* genome assembly, we sequenced the KitaakeX genome using a

78   strategy that combines short-read and long-read sequencing. Sequencing reads were collected using

79   Illumina, 10x Genomics, PACBIO, and Sanger platforms at the Joint Genome Institute (JGI) and the

80   HudsonAlpha Institute. The current release is version 3.0, which is a combination of a MECAT (Mapping,

81   Error Correction and de novo Assembly Tools) PACBIO based assembly and an Illumina sequenced 10x

82   genomics SuperNova assembly. The assembled sequence contains 377.6 Mb, consisting of 33 scaffolds

83   (476 contigs) with a contig N50 of 1.4 Mb, covering a total of 99.67% of assembled bases in chromosomes

84   (Table 1).

85   We assessed the quality of the KitaakeX assembly for sequence completeness and accuracy.

86   Completeness of the assembly was assessed by aligning the 34,651 annotated genes from the v7.0

87   Nipponbare to the KitaakeX assembly using BLAT [18]. The alignments indicate that 98.94% (34,285 of

88   genes) genes completely aligned to the KitaakeX assembly, 0.75% (259 genes) partially aligned, and 0.31%

89   (107 genes) were not detected. A bacterial artificial chromosome (BAC) library was constructed and a set

90   of 346 BAC clones (9.2x clone coverage) was sequenced using PACBIO sequencing. A range of variants

91   was detected by comparing the BAC clones to the assembly. Alignments were of high quality (<0.1% of

92   error) in 271 clones (Additional file 1: Figure S13). Sixty BACs indicate a higher error rate (0.45% of error)

93   due mainly to their placement in repetitive regions (Additional file 1: figure S14). Fifteen BAC clones indicate

94   a rearrangement (10 clones) or a putative overlap on adjacent contigs (5 clones) (Additional file 1: figure

95   S15). The overall error rate in the BAC clones is 0.09%, indicating the high quality of this assembly (for

96   detailed information, see Additional file 1).

97   We predicted 35,594 protein-coding genes in the KitaakeX genome (Table 1), representing 31.5%

98   genic space of the assembled genome size (Table 1). There is some transcriptome support for 89.5%

99   (31,854/35,594) of the KitaakeX genes, and 81.6% (29,039/35,594) genes are fully supported by the

100   transcriptome (Additional file 2 Table: S11). The predicted protein-coding genes are distributed unevenly

101   across each chromosome; gene density tends to be higher toward chromosome ends (Fig. 1i). The average

102   GC content of the genome is 43.7% (Fig. 1h, Table 1).

103   To assess the quality of the annotation of Kitaake genes, we compared the KitaakeX annotation to

104   those of other completed rice genomes using the BUSCO v2 method, which is based on a set of 1,440

105   conserved plant genes. The results confirm 99.0% completeness of the KitaakeX genome annotation

106   (Table1, Additional file 2: Table S7). To further evaluate the quality of the annotation, we studied the extent

107   of conservation of functional genes in KitaakeX. We selected 291 genes (Additional file 3) from three

108   pathways associated with stress resistance, flowering time and response to light [19], and then searched

109   for orthologous genes in the KitaakeX genome. We found that 275 of 291 (94.5%) of the selected KitaakeX

110   genes show greater than 90% identity with the corresponding Nipponbare genes at the protein level.

111   Twenty-three out of the 291 show 100% identity on the genome level but not on the protein level. Of these

112   23 genes, the KitaakeX gene model for 16 genes has better transcriptomic evidence than the Nipponbare

113   gene model. One of the 291 KitaakeX genes is slightly shorter than its Nipponbare ortholog due to an

114   alternative transcript (Additional file 3). These results indicate the high quality of the annotation, and

115   conservation between the KitaakeX and Nipponbare *japonica* rice varieties.

116   Using SynMap, we identified 2,469 pairs of colinear genes (88 blocks) in the KitaakeX genome

117   (Fig. 1j). These results correlate with already published findings [20]. We used RepeatMaker and Blaster

118   to identify transposable elements (TEs) in the KitaakeX genome, and identified 122.2 Mb of sequence

119   corresponding to TEs (32.0% of the genome). DNA transposons account for ~33 Mb; retrotransposons

120   account for ~90 Mb. The TEs belong mostly to the Gypsy and Copia retroelement families, and account for

121   23% of the genome (Additional file 2: Table S8), as is true in the Nipponbare and Zhenshan97 genomes

122   [21].

123   Table 1 Summary of the KitaakeX genome assembly and annotation

	Estimated genome size	409.5 Mb
	Assembled contigs	377.6 Mb
	Contig N50	1.4 Mb
Genome assembly	Longest contig	8.6 Mb
	Assembled scaffolds	381.6 Mb
	Scaffold N50	30.3 Mb
	Longest scaffold	44.3 Mb
	GC content	43.7%

Transposable elements	Retrotransposons	89.6 Mb
	DNA transposons	32.6 Mb
	Total	122.2 Mb
Genome annotation	Protein-coding genes	35,594
	Complete BUSCOs	99.0%
	Average transcript length	1,874 bp
	Average coding sequence length	1,222 bp
	Functionally annotated	33,226

124   We compared the genome of KitaakeX to the Nipponbare and Zhenshan97 genomes to detect

125   genomic variations, including single nucleotide polymorphisms (SNPs), insertions and deletions under 30

126   bp (InDels), presence/absence variations (PAVs), and inversions using MUMmer (Kurtz, Phillippy et al.

127   2004). We found 331,335 variations between KitaakeX and Nipponbare (Additional file 4), and nearly 10

128   times as many (2,785,991) variations between KitaakeX and Zhenshan97 (Additional file 5). There are

129   253,295 SNPs and 75,183 InDels between KitaakeX and Nipponbare, and 2,328,319 SNPs and 442,962

130   InDels between KitaakeX and Zhenshan97 (Additional files 6 and Additional file 2: Table S3). With respect

131   to SNPs in both intersubspecies (*japonica* vs. *indica*) as well as intrasubspecies (*japonica* vs. *japonica*)

132   comparisons, transitions (Tss) (G ->A and C ->T) are about twice as abundant as transversions (Tvs) (G -

133   >C and C ->G) (Additional file 2: Table S10). Genomic variations between KitaakeX and Nipponbare are

134   highly concentrated in some genomic regions (Fig. 1e), but variations between KitaakeX and Zhenshan97

135   are spread evenly through the genome (Fig. 1f). Intersubspecies genomic variations, then, are much more

136   extensive than intrasubspecies variations. We also detected multiple genomic inversions using comparative

137   genomics (Additional files 4 and 5).

138   For variations occurring in the genic regions, we found that single-base and 3 bp (without frame

139   shift) InDels are much more abundant than others (Additional file 7: Figure S16a), suggesting that these

140   genetic variations have been functionally selected. We carried out detailed analysis of gene structure

141   alterations that exist as a consequence of SNPs and InDels between KitaakeX and Nipponbare and Kitaake

142   and Zhenshan97. Between KitaakeX and Nipponbare, we identified 2,092 frameshifts, 78 changes affecting

6

143   splice-site acceptors, 71 changes affecting splice-site donors, 19 lost start codons, 161 gained stop codons,

144   and 15 lost stop codons. In the comparison of KitaakeX to Zhenshan97, 6,809 unique genes in KitaakeX

145   are affected by 8,640 frameshifts (Additional file 7: Figure S16b), 531 changes affecting splice-site

146   acceptors, 530 changes affecting splice-site donors, 185 lost start codons, 902 gained stop codons and

147   269 lost stop codons (Additional file 7: Figure S16b).

148   Based on PAV analysis, we identified 456 loci that are specific to KitaakeX (Additional file 4)

149   compared with Nipponbare. Pfam analysis of KitaakeX-specific regions revealed 275 proteins. Out of these

150   275 genes, 148 genes are from 19 different gene families with more than 2 genes in those regions. These

151   gene families include protein kinases, leucine-rich repeat proteins, NB-ARC domain-containing proteins, F-

152   box domain containing proteins, protein tyrosine kinases, Myb/SANt-like DNA binding domain proteins,

153   transferase family proteins, xylanase inhibitor C-terminal protein, and plant proteins of unknown function

154   (Additional file 7: Figure S16c). We identified 4589 loci specific to KitaakeX compared with Zhenshan97

155   (Additional file 5).

156   We also compared our *de novo* assembly of KitaakeX genome with Kitaake resequencing reads

157   using an established pipeline [22]. This analysis revealed 219 small variations (200 SNPs and 19 INDELs)

158   between the two genomes (Additional file 8). These variations affect 9 genes in KitaakeX besides the Ubi-

159   *Xa21* transgene, including the selectable marker encoding a hygromycin B phosphotransferase on

160   chromosome 6 (Additional file 8, Additional file 9: Figure S17).

161   **Discussion**

162   In 2005 the Nipponbare genome was sequenced and annotated to a high-quality level (International Rice

163   Genome Sequencing and Sasaki 2005). Since that time, it has served as a reference genome for many

164   rice genomic studies [23]. Despite its use, the long life cycle of Nipponbare makes it time-consuming for

165   most genetic analyses. Here we report the *de novo* assembly and annotation of KitaakeX, an early-flowering

166   rice variety with a rapid life cycle that is easy to propagate under greenhouse conditions. We predict that

167   KitaakeX contains 35,594 protein-coding genes, comparable to the published genomes (39,045 for

168   Nipponbare and 34,610 for Zhenshan97) (Additional file 4 and Additional file 5). The availability of a high-

7

169 quality genome and annotation for KitaakeX will be useful for associating traits of interest with genetic

170 variations, and for identifying the genes controlling those traits.

171 We identified 219 SNPs and InDels between the KitaakeX and Kitaake genomes. These variations

172 may have resulted from somatic mutations that arose during tissue culture and regeneration, or they may

173 be spontaneous mutations [24]. For rice, 150 mutations are typically induced during tissue culture and 41

174 mutations occur spontaneously per three generations [24]. These numbers are consistent with the

175 independent propagation of KitaakeX and Kitaake over approximately 10 generations in the greenhouse.

176 The KitaakeX genome will be useful for variety of studies. For example, we recently published the

177 whole genome sequences of 1,504 FN-mutated KitaakeX rice lines [22]. Mutations were identified by

178 aligning reads of the KitaakeX mutants to the Nipponbare reference genome [8]. On average, 97% of the

179 Nipponbare genome is covered by the KitaakeX reads. However, in some regions, the KitaakeX genome

180 diverges from Nipponbare to such an extent that no variants can be confidently identified. These appear

181 either as gaps in coverage or as regions containing a concentration of natural variations between KitaakeX

182 and Nipponbare. We can now use the KitaakeX sequence as the direct reference genome and detect

183 mutations in highly variable regions. This approach will simplify analysis and increase confidence in the

184 identification of FN-induced mutations.

185 ## Conclusions:

186 The *de novo* assembly of the KitaakeX genome serves as a useful reference genome for the model rice

187 variety Kitaake and will facilitate investigations into the genetic basis of diverse traits critical for rice biology

188 and genetic improvement.

189 ## Methods

190 **Plant Growth conditions**

191 Rice seeds were germinated on 1/2x MS (Murashige and Skoog) medium. Seedlings were transferred to

192 a greenhouse and planted 3 plants/pot during the springtime (Mar. 2, 2017) in Davis, California. The light

193 intensity was set at approximately 250 μmol m^{-2} s^{-1}. The day/night period was set to 14/10 h, and the

194 temperature was set between 28 and 30 °C [25]. Rice plants were grown in sandy soil supplemented with

8

195   nutrient water. The day when the first panicle of the plant emerged was recorded as the heading date for

196   that plant. Kasalath seeds were received later, and the heading date was recorded in the same way. The

197   experiment was repeated in winter.

198   **Construction of a phylogenetic tree**

199   We obtained 178,496 evenly distributed SNPs by dividing the genome into 3.8 kb bins and selecting one

200   or two SNPs per bin randomly according to the SNP density of the bin. Genotypes of all the rice

201   accessions, including 3,010 accessions of the 3K Rice Genomes Project and additional noted

202   accessions, were fetched from the SNP database RiceVarMap v2.0 [26] and related genomic data [27]

203   and used to calculate an IBS distance matrix which was then applied to construct a phylogenetic tree by

204   the unweighted neighbor-joining method, implemented in the R package APE [28]. Branches of the

205   phylogenetic tree were colored according to the classification of the 3,010 rice accessions [2].

206   **Genome Sequencing and Assembly**

207   High molecular weight DNA from young leaves of KitaakeX was isolated and used in sequencing.

208   See (Additional file 1) for further details.

209   **Annotation of Protein-Coding Genes**

210   To obtain high-quality annotations, we performed high throughput RNA-seq analysis of libraries from

211   diverse rice tissues (leaf, stem, panicle, and root). Approximately 683 million pairs of 2x151 paired-end

212   RNA-seq reads were obtained and assembled using a comprehensive pipeline PERTRAN (Shu,

213   unpublished). Gene models were predicted by combining *ab initio* gene prediction, protein-based

214   homology searches, experimentally cloned cDNAs/expressed-sequence tags (ESTs) and assembled

215   transcripts from the RNA-seq data. Gene functions were further annotated according to the best-matched

216   proteins from the SwissProt and TrEMBL databases [29] using BLASTP (E value < 10^{-5}) (Additional file

217   11). Genes without hits in these databases were annotated as "hypothetical proteins". Gene Ontology

218   (GO) [30] term assignments and protein domains and motifs were extracted with InterPro [31]. Pathway

9

219 analysis was derived from the best-match eukaryotic protein in the Kyoto encyclopedia of genes and

220 genomes (KEGG) database [32] using BLASTP (E value<$1.0e^{-10}$).

221 **Genome Synteny**

222 We used SynMap (CoGe, www.genomevolution.org) to identify collinearity blocks using homologous CDS

223 pairs with parameters according to Daccord et al [33] and visualized collinearity blocks using Circos [34].

224 **Repeat Annotation**

225 The fraction of transposable elements and repeated sequences in the assembly was obtained merging

226 the output of RepeatMasker (http://www.repeatmasker.org/, v. 3.3.0) and Blaster (a component of the

227 REPET package) [35]. The two programs were run using nucleotide libraries (PReDa and

228 RepeatExplorer) from RiTE-db [36] and an in-house curated collection of transposable element (TE)

229 proteins, respectively. Reconciliation of masked repeats was carried out using custom Perl scripts and

230 formatted in gff3 files. Infernal [37] was adopted to identify non-coding RNAs (ncRNAs) using the Rfam

231 library Rfam.cm.12.2 [38]. Results with scores lower than the family-specific gathering threshold were

232 removed; when loci on both strands were predicted, only the hit with the highest score was kept. Transfer

233 RNAs were also predicted using tRNAscan-SE [39] at default parameters. Repeat density was calculated

234 from the file that contains the reconciled annotation (Additional file 10).

235 **Analysis of Genomic Variations**

236 Analysis of SNPs and InDels: We used MUMmer (version 3.23) [40] to align the Nipponbare and

237 Zhenshan97 genomes to the KitaakeX genome using parameters -maxmatch -c 90 -l 40. To filter the

238 alignment results, we used the delta -filter -1 parameter with the one-to-one alignment block option. To

239 identify SNPs and InDels we used show-snp option with parameter (-Clr TH). We used snpEff [41] to

240 annotate the effects of SNPs and InDels. Distribution of SNPs and InDels along the KitaakeX genome

241 was visualized using Circos [34].

242   Analysis of PAVs and inversions: We used the show-coords option of MUMmer (version 3.23)

243 with parameters -TrHcl to identify gap regions and PAVs above >86 bp in size from the alignment blocks.

10

244   We used the inverted alignment blocks with ≥98% identity from the show-coords output file to identify

245   inversions.

246   To identify genomic variations between Kitaake and KitaakeX we sequenced and compared the

247   sequences using the established pipeline [22].

248   **BAC library construction**

249   Arrayed BAC libraries were constructed using established protocols [42]. Please see Additional

250   file 1 for further details.

251

252

253   **Additional files:**

254   Additional file 1: Supplemental method: Tables S1-S6, Figures S1-S15

255   Additional file 2: Comparison of the KitaakeX genome with other rice genomes and KitaakeX annotation;

256   Tables S7-S11

257   Additional file 3: Genes used in annotation quality control

258   Additional file 4: Comparative genomics between KitaakeX and Nipponbare

259   Additional file 5: Comparative genomics between KitaakeX and Zhenshan97

260   Additional file 6: SNPs between KitaakeX and Zhenshan97

261   Additional file 7: Figure S16. Genomic variations showing gene variations between KitaakeX and

262   Nipponbare and Zhenshan97

263   Additional file 8. Genomic variations between KitaakeX and Kitaake

264   Additional file 9: Figure S17. Position of the XA21 locus in the KitaakeX genome

265   Additional file 10: Repeat annotation

266   Additional file 11: Gene functional annotation

11

270   for providing the Kitaake seeds to PCR in 1995. These seeds were provided to Dr. Okita by Dr. Hiroyuki

271   Ito, Akita National College of Technology, Japan. We thank Dr. Jan E. Leah at Colorado State University

272   for seeds of Zhenshan 97, Minghui 63, IR64 and 93-11, and the USDA Dale Bumpers National Rice

273   Research Center, Stuttgart, Arkansas for seeds of Kasalath.

281   **Availability of data and material:**

282   The genome sequencing reads and assembly have been deposited under GenBank under accession

283   number PRJNA234782 and PRJNA448171 respectively. The assembly and annotation of the Kitaake

284   genome are available at Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). The RNA-Seq reads

285   of KitaakeX leaf, panicle, stem and root have been deposited under GenBank accession numbers

286   SRP182736, SRP182738, SRP182741, and SRP182737 respectively. Genome sequencing reads for

287   Kitaake have been deposited under GenBank under accession number SRP193308.

288   **AUTHOR CONTRIBUTIONS**

289   RJ, G.L, M.C.R. and P.C.R conceived and initiated the study. R.J. and G.L carried out sequencing,

290   assembly and annotation in collaboration with J.J., S.S., D.A.K., J.T., D.D., D.B., J.G., D.C., K.W.B.,

291   R.A.W., N.T.P, and J.S. R.J. and G.L conducted comparative genomics. F.X and W.X contributed to

292   phylogenetic tree. R.J., P.C.R, M.S C, G.L and C.R.N. wrote the manuscript. All authors read and

293   approved the final manuscript.

294 **Ethics approval and consent to participate**

295 Not applicable

296 **Consent to Publication**

297 Not applicable

298 **Competing interests:**

299 The authors declare no conflict of interest.

300 **Fig. 1 The early flowering rice variety KitaakeX.**

301 **a** KitaakeX and selected sequenced rice varieties under long-day conditions. Scale bar = 10 cm; **b**

302 Flowering time of KitaakeX and selected rice varieties under long-day conditions. DAG, days after

303 germination. Asterisks indicate significant differences using the unpaired Student's *t*-test (P < 0.0001); **c**

304 KitaakeX in the unweighted neighbor-joining tree comprising 3,010 accessions of the 3k rice genomes

305 project and indicated varieties. It includes four XI clusters (XI-1A from East Asia, XI-1B of modern varieties

306 of diverse origins, XI-2 from South Asia and XI-3 from Southeast Asia); three GJ clusters [primarily East

307 Asian temperate (named GJ-tmp), Southeast Asian subtropical (named GJ-sbtrp) and Southeast Asian

308 Tropical (named GJ-trp)]; and two groups for the mostly South Asian cA (circum-Aus) and cB (circum-

309 Basmati) accessions, 1 group Admix (accessions that fall between major groups were classified as

310 admixed) Branch length indicates the genetic distance between two haplotype**s; d** Circles indicate the 12

311 KitaakeX chromosomes represented on a Mb scale; **e,f** SNPs and InDels between KitaakeX and

312 Nipponbare (**e**) and Kitaake and   Zhenshan97 (**f**); **g** Repeat density; **h** GC content; **i** Gene density; **j**

313 Homologous genes in the KitaakeX genome. Window size used in the circles is 500 kb.

314

315

316

317   References for Section A
318

13

319   1.   Gross BL, Zhao Z: **Archaeological and genetic insights into the origins of**
320   **domesticated rice.** 2014, **111:**6190-6197.
321   2.   Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR,
322   Zhang FJN: **Genomic variation in 3,010 diverse accessions of Asian cultivated rice.**
323   2018, **557:**43.
324   3.   lchitani K, Okumoto Y, Tanisaka T: **Photoperiod sensitivity gene of se-1 locus found in**
325   **photoperiod insensitive rice cultivars of the northern limit region of rice cultivation.**
326   *Breeding science* 1997, **47:**8.
327   4.   Kim SL, Choi M, Jung KH, An G: **Analysis of the early-flowering mechanisms and**
328   **generation of T-DNA tagging lines in Kitaake, a model rice cultivar.** *J Exp Bot* 2013,
329   **64:**4169-4182.
330   5.   Kunihiro Y, Ebe Y, Wada S, Shinbashi N, Honma A, Sasaki T, Sasaki K, Numao Y,
331   Morimura K, Tan No H: **The new rice variety Kita-ake.** *Bulletin of Hokkaido*
332   *prefectural agricultural experiment stations* 1989, **59:**4.
333   6.   Wang L, Zheng J, Luo Y, Xu T, Zhang Q, Zhang L, Xu M, Wan J, Wang MB, Zhang
334   CJPbj: **Construction of a genomewide RNA i mutant library in rice.** 2013, **11:**997-
335   1005.
336   7.   Gao H, Zheng XM, Fei G, Chen J, Jin M, Ren Y, Wu W, Zhou K, Sheng P, Zhou F, et al:
337   **Ehd4 encodes a novel and Oryza-genus-specific regulator of photoperiodic**
338   **flowering in rice.** *PLoS Genet* 2013, **9:**e1003281.
339   8.   Li G, Jain R, Chern M, Pham NT, Martin JA, Wei T, Schackwitz WS, Lipzen AM,
340   Duong PQ, Jones KC, et al: **The Sequences of 1504 Mutants in the Model Rice**
341   **Variety Kitaake Facilitate Rapid Functional Genomic Studies.** *Plant Cell* 2017,
342   **29:**1218-1231.
343   9.   Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai
344   WX, Zhu LH, et al: **A receptor kinase-like protein encoded by the rice disease**
345   **resistance gene, Xa21.** *Science* 1995, **270:**1804-1806.
346   10.   Gao H, Jin M, Zheng X-M, Chen J, Yuan D, Xin Y, Wang M, Huang D, Zhang Z, Zhou
347   KJPotNAoS: **Days to heading 7, a major quantitative locus determining photoperiod**
348   **sensitivity and regional adaptation in rice.** 2014, **111:**16337-16342.
349   11.   Ronald PC, Beutler B: **Plant and animal sensors of conserved microbial signatures.**
350   *Science* 2010, **330:**1061-1064.
351   12.   Liu Y, Wu H, Chen H, Liu Y, He J, Kang H, Sun Z, Pan G, Wang Q, Hu JJNb: **A gene**
352   **cluster encoding lectin receptor kinases confers broad-spectrum and durable insect**
353   **resistance in rice.** 2015, **33:**301.
354   13.   Zhou X, Liao H, Chern M, Yin J, Chen Y, Wang J, Zhu X, Chen Z, Yuan C, Zhao
355   WJPotNAoS: **Loss of function of a rice TPR-domain RNA-binding protein confers**
356   **broad-spectrum disease resistance.** 2018**:**201705927.
357   14.   Rodrigues JA, Ruan R, Nishimura T, Sharma MK, Sharma R, Ronald PC, Fischer RL,
358   Zilberman D: **Imprinted expression of genes and small RNA is associated with**
359   **localized hypomethylation of the maternal genome in rice endosperm.** *Proc Natl*
360   *Acad Sci U S A* 2013, **110:**7934-7939.
361   15.   Li T, Liu B, Spalding MH, Weeks DP, Yang B: **High-efficiency TALEN-based gene**
362   **editing produces disease-resistant rice.** *Nat Biotechnol* 2012, **30:**390-392.

14

363 16. Xie K, Minkenberg B, Yang Y: **Boosting CRISPR/Cas9 multiplex editing capability**
364   **with the endogenous tRNA-processing system.** *Proc Natl Acad Sci U S A* 2015,
365   **112:**3570-3575.
366 17. Jung K-H, An G, Ronald PC: **Towards a better bowl of rice: assigning function to tens**
367   **of thousands of rice genes.** *Nature Reviews Genetics* 2008, **9:**91.
368 18. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
369 19. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, Zuccolo A, Song X,
370   Kudrna D, Ammiraju JS, et al: **The genome sequence of African rice (Oryza**
371   **glaberrima) and evidence for independent domestication.** *Nat Genet* 2014, **46:**982-
372   988.
373 20. Guyot R, Keller BJG: **Ancestral genome duplication in rice.** 2004, **47:**610-614.
374 21. Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song J-M, Xie
375   W, et al: **Extensive sequence divergence between the reference genomes of two elite**
376   **indica rice varieties Zhenshan 97 and Minghui 63.** 2016, **113:**E5163-
377   E5171.
378 22. Li G, Jain R, Chern M, Pham NT, Martin JA, Wei T, Schackwitz WS, Lipzen AM,
379   Duong PQ, Jones KCJTPC: **The sequences of 1,504 mutants in the model rice variety**
380   **Kitaake facilitate rapid functional genomic studies.** 2017:tpc. 00154.02017.
381 23. Matsumoto T, Wu J, Itoh T, Numa H, Antonio B, Sasaki T: **The Nipponbare genome**
382   **and the next-generation of rice genomics research in Japan.** *Rice (New York, NY)*
383   2016, **9:**33-33.
384 24. Tang X, Liu G, Zhou J, Ren Q, You Q, Tian L, Xin X, Zhong Z, Liu B, Zheng XJGb: **A**
385   **large-scale whole-genome sequencing analysis reveals highly specific genome editing**
386   **by both Cas9 and Cpf1 (Cas12a) nucleases in rice.** 2018, **19:**84.
387 25. Schwessinger B, Bahar O, Thomas N, Holton N, Nekrasov V, Ruan D, Canlas PE, Daudi
388   A, Petzold CJ, Singan VR, et al: **Transgenic expression of the dicotyledonous pattern**
389   **recognition receptor EFR in rice leads to ligand-dependent activation of defense**
390   **responses.** *PLoS Pathog* 2015, **11:**e1004809.
391 26. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, Xing Y, Chen L, Xie W:
392   **RiceVarMap: a comprehensive database of rice genomic variations.** *Nucleic Acids*
393   *Res* 2015, **43:**D1018-1022.
394 27. Li G, Chern M, Jain R, Martin JA, Schackwitz WS, Jiang L, Vega-Sanchez ME, Lipzen
395   AM, Barry KW, Schmutz J, Ronald PC: **Genome-Wide Sequencing of 41 Rice (Oryza**
396   **sativa L.) Mutated Lines Reveals Diverse Mutations Induced by Fast-Neutron**
397   **Irradiation.** *Mol Plant* 2016, **9:**1078-1081.
398 28. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R**
399   **language.** *Bioinformatics* 2004, **20:**289-290.
400 29. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its**
401   **supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
402 30. Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T:
403   **A gene ontology inferred from molecular networks.** *Nat Biotechnol* 2013, **31:**38-45.
404 31. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY,
405   Dosztanyi Z, El-Gebali S, Fraser M, et al: **InterPro in 2017-beyond protein family and**
406   **domain annotations.** *Nucleic Acids Res* 2017, **45:**D190-D199.
407 32. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives**
408   **on genomes, pathways, diseases and drugs.** *Nucleic Acids Res* 2017, **45:**D353-D361.

409 33. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H,
410   Bianco L, Micheletti D, Velasco R, et al: **High-quality de novo assembly of the apple**
411   **genome and methylome dynamics of early fruit development.** *Nat Genet* 2017,
412   **49:**1099-1106.
413 34. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra
414   MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009,
415   **19:**1639-1645.
416 35. Flutre T, Duprat E, Feuillet C, Quesneville H: **Considering transposable element**
417   **diversification in de novo annotation approaches.** *PLoS One* 2011, **6:**e16526.
418 36. Copetti D, Zhang J, El Baidouri M, Gao D, Wang J, Barghini E, Cossu RM, Angelova A,
419   Maldonado LC, Roffler S, et al: **RiTE database: a resource database for genus-wide**
420   **rice genomics and evolutionary biology.** *BMC Genomics* 2015, **16:**538.
421 37. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.**
422   *Bioinformatics* 2013, **29:**2933-2935.
423 38. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW,
424   Gardner PP, Jones TA, Tate J, Finn RD: **Rfam 12.0: updates to the RNA families**
425   **database.** *Nucleic Acids Res* 2015, **43:**D130-137.
426 39. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web**
427   **servers for the detection of tRNAs and snoRNAs.** *Nucleic Acids Res* 2005, **33:**W686-
428   689.
429 40. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg
430   SLJGb: **Versatile and open software for comparing large genomes.** 2004, **5:**R12.
431 41. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden
432   DM: **A program for annotating and predicting the effects of single nucleotide**
433   **polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain**
434   **w1118; iso-2; iso-3.** *Fly (Austin)* 2012, **6:**80-92.
435 42. Luo M, Wing RA: **An Improved Method for Plant BAC Library Construction.** In
436   *Plant Functional Genomics.* Edited by Grotewold E. Totowa, NJ: Humana Press; 2003:
437   3-19
438

a

KitaakeX
Nipponbare 93-11 IR64 Zhenshan 97 Minghui 63 Kasalath

b

Heading date (DAG)

KitaakeX Nipponbare 93-11 IR64 Zhenshan 97 Minghui 63 Kasalath

c

KitaakeX Nipponbare

Minghui 63

93-11

IR64

Kasalath

Zhenshan 97

0.05

| GJ-tmp | GJ-sbtrp | GJ-trp | cA (Aus) | cB (Bas) |
| XI-1A | XI-1B | XI-2 | XI-3 | Admix |

d CHR12 CHR1
e
f
g
h
i
j

CHR11
CHR10
CHR9
CHR8
CHR7 CHR6 CHR5
CHR2
CHR3
CHR4