SnpReady for Rice (SR4R) Database

Jun Yan[1,a#], Dong Zou[2,b#], Chen Li[3,c], Zhang Zhang[2,d], Shuhui Song[2,e*], Xiangfeng Wang[1,f*]

[1] *Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China*

[2] *National Genomics Data Center & BIG Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Bejing 100101, China*

[3] *Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China*

[#] Equal contribution

[*] Correspondence should be addressed to **xwang@cau.edu.cn (Wang XF) and songshh@big.ac.cn (Song SH);**

Emails of authors:

hh405262448@126.com (Yan J);

zoud@big.ac.cn (Zou D);

lic11111@sina.com (Li C);

zhangzhang@big.ac.cn (Zhang Z)

[a] ORCID: 0000-0002-3806-6457

[b] ORCID: 0000-0002-7169-4965

[c] ORCID: 0000-0001-6702-6860

[d] ORCID: 0000-0001-6603-5060

26 e ORCID: 0000-0003-2409-8770

27 f ORCID: 0000-0002-6406-5597

28

29 Number of words: 5421

30 Number of figures: 7

31 Number of tables: 0

32 Number of supplementary figures: 4

33 Number of supplementary tables: 3

34

35 **Running title:** *Jun Yan et al. / SNP Ready for Rice database*

36

37

38

39

Abstract

The information commons for rice (IC4R) database is a collection of ~18 million SNPs (single nucleotide polymorphisms) identified by the resequencing of 5,152 rice accessions. Although IC4R offers ultra-high density rice variation map, these raw SNPs are not readily usable for the public. To satisfy different research utilizations of SNPs for population genetics, evolutionary analysis, association studies and genomic breeding in rice, the raw genotypic data of the 18 million SNPs were processed by unified bioinformatics pipelines. The outcomes were used to develop a daughter database of IC4R – SnpReady for Rice (SR4R). The SR4R presents four reference SNP panels, including 2,097,405 hapmapSNPs after data filtration and genotype imputation, 156,502 tagSNPs selected from linkage disequilibrium (LD)-based redundancy removal, 1,180 fixedSNPs selected from genes exhibiting selective sweep signatures, and 38 barcodeSNPs selected from DNA fingerprinting simulation. SR4R thus offers a highly efficient rice variation map that combines reduced SNP redundancy with extensive data describing the genetic diversity of rice populations. In addition, SR4R provides rice researchers with a web-interface that enables them to browse all four SNP panels, use online toolkits, and retrieve the original data and scripts for a variety of population genetics analyses on local computers. The SR4R is freely available to academic users at http://sr4r.ic4r.org/.

Keywords: Rice; SNP; Database; Hapmap

Introduction

Oryza sativa, or rice, was the first crop genome to be sequenced. In the past decade, thousands of rice accessions in the germplasm banks worldwide have been genotyped [1] and numerous rice variation databases have been constructed. One of these databases is the rice variation database (RVD), a daughter database of the Information Commons for Rice consortium (IC4R) [2]. RVD is a collection of over eighteen million SNPs (single nucleotide polymorphisms) identified from 5,152 rice accessions based on whole-genome resequencing data, and offers an ultra-high-density rice variation map – about one SNP per twenty bases on average. The information contained in this high volume of raw SNPs is not ready for use until

70   it has been processed to remove low-quality SNPs, such as those with missing/low frequency

71   genotypes, or redundant SNPs identified due to linkage disequilibrium (LD). In addition,

72   different types of research require different magnitudes of SNPs to ensure efficient

73   computing and accurate results; for example, the requirements are different for evolutionary

74   studies using comparative genomics and pan-genome analysis, gene mapping by quantitative

75   trait loci (QTL), genome-wide association study analysis (GWAS), molecular breeding by

76   marker-assisted selection (MAS) and genomic selection (GS), and variety protection by DNA

77   fingerprint barcoding.

78   Construction of a reference haplotype map (HapMap) to represent the maximal population

79   diversity for a species is the first step. The ~18 million raw SNPs in RVD provide an initial

80   variation dataset to generate a reference HapMap for rice. According to international human

81   HapMap database, which contains over 3.1 million high-quality SNPs, a density of one SNP

82   per 100 bases is sufficient for performing genotype imputation, GWAS analysis and mapping

83   of causal variations [3]. Because the genome size of rice is ~ 400 Mbp, about two million

84   high-quality SNPs may offer an ideal density of one SNP per 200 bases. Such density of a

85   reference rice HapMap is especially useful for molecular breeders to perform genotype

86   imputation to supplement missing genotypes or increase SNP density, as low-density

87   genotyping platforms are mostly used in rice to lower genotyping expense.

88   For population genetics studies in which thousands of individual samples are assessed, the

89   millions of SNPs in an entire HapMap are excessive. The redundant SNPs in a HapMap

90   extensively increase computing costs, and may also reduce the accuracy of results. To

91   circumvent these challenges, a subgroup of SNPs whose genotypes significantly correlate

92   with other SNPs in the same linkage disequilibrium (LD) region are selected; these are

93   known as tagging SNPs. The number of tagging SNPs may vary between species and

94   populations, depending upon the lengths of LD regions in each group [4]. Based on the data

95   in RVD, LD length in rice ranges from 100 to 500 Kb; thus 100,000 SNPs, which yields a

96   density of one tagging SNP per 3 to 5 Kbp, is sufficient for various genetic diversity analysis.

97   The expense of genotyping is an important factor to consider in crop molecular breeding,

98   as molecular breeding typically requires the rapid genotyping of thousands of samples, often

99   within days or even hours. Therefore, low SNP density genotyping technologies, such as SNP

100 chip or KASP-based platforms are usually preferred by industrial seed companies; these

101 methods offer great flexibility by combining the rapid identification of low numbers of SNPs

102   (several to a few dozen) with the ability to multiplex hundreds to thousands of DNA samples.

103   However, these methods lack precision.

104   Modern breeding methods demand the efficiency and stability of a highly concise marker

105   panel containing ~1K SNPs. SNPs used to select plants for breeding typically occur in genes

106   or genomic regions that are associated with agronomic traits believed to be subject to

107   selective pressures [5]. Genes with variations exhibiting selectively fixed signatures can be

108   identified based on the $\theta\pi$ and Fst values computed by selective sweep analysis [6]. This

109   magnitude of SNPs is suitable for synthesis on low-density SNP chips, which are then used

110   for conducting certain types of molecular analysis, such as marker-assisted selection, seed

111   purity or heterozygosity testing, genetic component analysis, and subpopulation classification.

112   For intellectual protection of commercial rice varieties, DNA fingerprinting typically uses

113   only 12 to 36 SNPs, to generate a combination of barcodes with maximal resolution to

114   distinguish commercial varieties in the seed industry or germplasm accessions in gene banks.

115   Simulation of all possible combinations of a set of candidate SNPs have to be tested in a large

116   germplasm population to ensure the maximal resolution with fewest markers, such as the

117   MinimalMarker algorithm [7].

118   To enhance the ability of researchers to effectively use the RVD in IC4R, we developed a

119   daughter database we have called SnpReady for Rice, or SR4R. SR4R enables researcher to

120   readily retrieve SNPs that are relevant to their own research, thus saving time and

121   computational resources. In SR4R, the ~18 million SNPs have been divided into four

122   categories: hapmapSNPs, tagSNPs, fixedSNPs, and barcodeSNPs (**Figure 1**). SR4R allows

123   users to browse the related information associated with each SNP panel, and also to

124   download each set of genotype files for local use. SR4R also offers 18 bioinformatics tools

125   and pipeline scripts, enabling users to locally run the tools to perform genotype imputation,

126   basic statistical analysis, genotype file format conversion, SNPs filtration and extraction,

127   population structure analysis, genetic diversity analysis, rice subpopulation classification,

128   DNA fingerprinting analysis, and other additional functions.

129

130   Database contents and analytical modules

131   The hapmapSNP panel

132   The IC4R rice variation database (RVD; http://variation.ic4r.org/) is a collection of over 18

133   million SNPs with related annotation information, identified from previously published

134   whole genome resequencing of 5,152 rice accessions [2]. Such a high-density rice variation

135   map, which identifies an average of one SNP per twenty bases, offers the possibility of

136   generating a high-density HapMap for the rice research community; creating such a HapMap

137   was the first step in creating the SnpReady for Rice (SR4R) Database described here.

138   To ensure the quality of HapMap, we performed an initial filtration of samples and SNPs

139   on the raw dataset of 5,152 accessions (**Materials and Methods**). First, a total of 2,556

140   accessions with genotype missing rate less than 20% were selected; each selected accession

141   has been documented with explicit subpopulation classification and origins (**Table S1**). Then,

142   SNPs with genotype missing rate $\geq$ 0.1 and minor allele frequency (MAF) $\leq$ 0.05 were

143   removed. Genotype imputation on the resulting 2,883,623 SNPs in the selected 2,556

144   accessions yielded a high-quality HapMap containing 2,097,405 SNPs without any missing

145   genotypes using the software Beagle [8]. These 2,097,405 SNPs were regarded as the

146   hapmapSNP panel, and were used as the initial dataset for generating the other three SNP

147   panels (**Figure 2A and 2D**).

148   The generated reference HapMap of rice has an average density of five SNPs per Kb with

149   a heterozygosity rate of 3.75% (**Figure 2B and 2D**). Genome-wide distribution statistics

150   showed that 58.4% of the hapmapSNPs present in the intergenic regions, 12.5% in the

151   intronic regions, 11.8% in the exonic regions, 0.02% on the splicing sites, and 10.6% and

152   6.8% hapmapSNPs located in the upstream and downstream regions (1Kb away from

153   transcription start site or transcription end site) of a gene territory (**Figure 2F**). The 2,097,405

154   hapmapSNPs with genotypes of 2,556 accessions are available to download, enabling users to

155   perform genotype imputation on local genotype data to increase the density of SNPs

156   generated from low-density genotyping platform.

157

158   **The tagSNP panel**

159   High SNP density is usually beneficial to precise mapping of trait-related genes with GWAS

160   analysis, but is not suitable for population genetic analysis because SNP redundancy may add

161   unnecessary computation costs and introduce bias to the results [9]. Since SNPs within the

162   same LD region possess correlated genotypes forming one haplotype block, a representative

163   SNP is usually selected as a tag to solve the redundancy issue. We adopted an LD-based SNP

164   pruning procedure to infer haplotype tagging SNPs (tagSNPs) from the hapmapSNPs

165   (**Materials and Methods**). As a result, 156,502 tagSNPs were identified **(Figure 1).** To

166   verify whether the tagSNP panel properly represents the genetic diversity of the population,

167   phylogenetic analysis using the 156,502 tagSNPs was performed on the 2,556 rice accessions

168   which were explicitly documented with subpopulation classification and origins. As shown in

169   **Figure 3A**, the resulting phylogenetic tree clearly exhibited six major clades representing the

170   five cultivated rice subpopulations and one wild rice subpopulation. The five cultivated rice

171   subpopulations include *indica* rice (*Ind* for short) containing 1,655 accessions, *Aus* rice (*Aus*)

172   containing 182 accessions, *Aromatic* (*Aro*) rice containing 56 accessions, tropical *japonica*

173   rice (*TrJ*) containing 318 accessions, and temperate *japonica* rice (*Tej*) containing 327

174   accessions, whilst the wild rice subpopulation contains 18 *O. rufipogon* (*Oru*) accessions. In

175   addition, PCA-based (**Figure 3B**) and admixture-based (**Figure 3C**) analyses showed the

176   same pattern, with the subpopulation classification as the phylogenetic tree indicated. For

177   population admixture structure analysis, a predefined parameter of "K value" was used to

178   mandatorily estimate the number of ancestral subpopulation and uses different colours for

179   each K value to represent the number of subpopulations. Because the optimal number of

180   ancestral subpopulation is usually unknown, a common way is to use a series of K value to

181   estimate the optimal K parameter. It is worth noting that the *japonica*, *indica* and *Aus*

182   subpopulations were explicitly separated when K was set to 3, while the six subpopulations

183   were clearly separated until the K value was set to 8. In addition, between K=4 to 7, the

184   *indica* subpopulation showed clear structure divided into six groups (*indica* g1 to g6) as

185   indicated by both PCA and admixture analysis (**Figure 3D and Figure S1**). The genetic

186   structures of the six rice subpopulations and the six *indica* subgroups are consistent with

187   multiple previous reports [10].

188

189   **Genetic diversity analysis with the tagSNP panel**

190   The tagSNP panel represents a subset of the hapmapSNPs after approximately 92.5% of the

191   genetic redundancy was removed (**Figure 1**). To test the effectiveness of the 156,502

192   tagSNPs, we performed another series of standard genetic diversity analyses and examined

193   whether the results agreed with previously reported conclusions. First, we found that the

194   count of homozygous SNPs and the heterozygosity rate of the accessions in the six

195 subpopulations showed opposite trends: while the accessions in the *TeJ* subpopulation had
196 the highest count of homozygous SNPs and lowest heterozygous rate, the accessions in the
197 *indica* subpopulation had the lowest count of homozygosity SNPs and highest homozygosity
198 rate (**Figure 4A and 4B**). The IBS (identity by state) analysis is a commonly used method to
199 measure the similarity of alleles in a designated population, which may reflect the genetic
200 diversity of the whole population and subpopulations. Comparison of the IBS values among
201 different subpopulations may help understand the degree of genetic differentiation in
202 different subpopulations. In order to validate whether the IBS results generated from the
203 tagSNPs are consistent with the previous reports regarding the genetic diversity in different
204 subpopulations, pairwise computation of the IBS values between each pair of accessions
205 within the same subpopulation was performed, and the results showed that temperate
206 *japonica* rice has the highest IBS values, while the *indica* rice has the lowest (**Figure 4C**). In
207 addition, runs of homozygosity (ROH) analysis indicated that the temperate *japonica* rice has
208 the most and longest ROH regions, while the *indica* rice has the least and shortest ROH
209 regions (**Figure 4D**). This pattern agreed with the result from LD decay analysis showing that
210 temperate *japonica* rice has the slowest LD decay rate while the *indica* rice has fastest rate
211 (**Figure 4E**). Computations of $\theta\pi$ and *Fst* are commonly used methods to measure genetic
212 diversity within population and between population, respectively (**Materials and Methods**).
213 The within-subpopulation diversities of the six rice subpopulations are *Oru* ($\theta\pi$=0.218), *Ind*
214 ($\theta\pi$=0.216), *Aus* ($\theta\pi$=0.182), *Aro* ($\theta\pi$=0.145), *TrJ* ($\theta\pi$=0.116) and *TeJ* ($\theta\pi$=0.068). Using the
215 wild rice subpopulation as reference, the genetic distances of the five types of cultivated rice
216 between wild rice are *TeJ* (*Fst*=0.476), *TrJ* (*Fst*=0.419), *Aus* (*Fst*=0.299), *Ind* (*Fst*=0.266)
217 and *Aro* (*Fst*=0.241), suggesting the highest domestication level of *japonica* rice compared to
218 other rice (**Figure 4F and 4G**). The collective results from multiple angles of standard
219 genetic diversity analyses were consistent with previous reports that *indica* rice has a more
220 complicated genetic composition and origin compared to the other five subpopulations [11].

221

Genomic selection analysis with the tagSNP panel

223 Genomic selection (GS) has been widely used in industrial animal and crop breeding
224 programs [12]. GS is essentially a best linear unbiased prediction (BLUP) model that is first
225 trained with known genotypes and phenotypes of reference population individuals, usually
226 accounting for 20% to 50% of a breeding population, and then used to predict the unknown

227 phenotypes of the remaining genotyped individuals (the candidate population). The predicted

228 phenotypes, known as the genomic estimated breeding values (GEBV), are ranked from high

229 to low, and can be used to assist in deciding upon a hybridization plan. Although GS may

230 significantly shorten the breeding cycle, the cost for genotyping has been a vital factor

231 because the GS model has to take genome-wide SNP markers as input, especially from crop

232 breeding in which thousands to hundreds of thousands of individuals need to be genotyped.

233 In order to lower genotyping cost, compilation of a set of thousands of SNPs that may best

234 represent the overall genetic backgrounds of a breeding population is of great importance.

235 Because the 156,502 tagSNPs category is a high-quality marker set with most redundancy

236 removed while preserving maximal genetic diversity, it may be considered as a marker pool

237 for selecting high-efficiency SNPs for genomic selection. To test the effectiveness, we

238 analysed a previously published dataset containing 414 rice parental lines with non-missing

239 genotypes of 29,434 SNPs profiled by the 44K rice SNP chip, and nine phenotype traits

240 (flowering time, panicle fertility, seed width, seed volume, seed surface area, plant height,

241 flag leaf length, flag leaf width, and florets per panicle). The GS model was obtained from

242 the ridge regression best linear unbiased prediction (rrBLUP) algorithm [13], and prediction

243 accuracy was evaluated with Pearson correlation between observed and predicted traits by

244 five-fold cross validation. The evaluation was performed using five different SNP

245 combinations: Set-1, the original 29,434 SNPs on the 44K chip; Set-2, the 1,090 SNPs

246 overlapped between the 156,502 tagSNPs and 29,434 SNPs; Set-3, the 1,090 SNPs randomly

247 selected from the 29,434 SNPs; Set-4, the 1,090 SNPs evenly distributed in the genome (350

248 Kb per SNP) selected from the 29,434 SNPs; and Set-5, the 1,090 consecutive SNPs

249 localized within a randomly selected genomic region from the 29,434 SNPs. Then the

250 rrBLUP prediction was performed on the nine phenotypes using the five sets of SNPs to

251 compare prediction accuracies (**Figure 5**). Although prediction accuracies greatly varied

252 ranging from 0.23 to 0.90 among the nine traits due to different heritability of each trait, the

253 trend of the five SNP sets within the same trait was generally consistent. Except for the trait

254 of panicle fertility in which the Set-2 SNPs exhibited the highest prediction accuracy, the full

255 29,434 SNPs showed the highest prediction accuracy for the other eight traits followed by the

256 1,090 tagSNPs in the second position. We further performed pairwise student's t-test for

257 Pearson correlations of the selected 1,090 tagSNPs set (Set-2) and other four sets, the result

258 shows that the selected 1,090 tagSNPs set significantly outperform other randomly selected

259 SNP set for most traits (**Figure S2**). These results indicate that selection of about one

260   thousand tagSNPs from the tagSNP pool might be a feasible option to lower genotyping

261   budget; for example, these SNPs could inform the synthesis of a new low-density SNP chip

262   rather than using high-density SNP chip.

263

264   **The fixedSNP panel**

265   In the crop breeding industry, genotyping cost-per-sample is a top-priority factor, since

266   hundreds to thousands of samples are often genotyped in single day. The data then assists a

267   variety of molecular breeding practices, including genomic selection-assisted phenotype

268   prediction, marker-assisted backcrossing, seed purity or genotype heterozygosity analysis,

269   and subpopulation identification. Cost reduction is usually fulfilled by compiling a highly

270   effective marker panel containing only dozens to hundreds of SNPs that are available for

271   high-throughput genotyping platforms, such as Douglas ArrayTape and LGC Omega-F

272   equipment, using a PCR-based KASP™ genotyping assay. These systems allow users to

273   flexibly combine different numbers of SNPs and DNA samples using multiple plates with 96

274   and 384 wells per run. To meet the industrial demand, further compression of the tagSNP

275   panel must consider not only the genetic relationship between subpopulations and accessions,

276   but also the evolutionary and/or functional significance of SNPs with high diagnostic

277   effectiveness and stability.

278   The *Fst* and *θπ* values are commonly used indicators of genomic regions demonstrating

279   signatures of selective sweeps, caused by domestications, artificial selections and

280   environmental adaption. SNPs in selective-sweep regions are usually evolutionarily fixed

281   with strong positive selection signals. To generate the fixedSNP panel, we first identified the

282   selective sweep regions that are specific to each subpopulation and are common to the six

283   subpopulations by combining the ratio of *Fst* versus *θπ* based on the comparison of the

284   cultivated subpopulation against the wild rice population (**Materials and Methods**). Using

285   100 Kb and 10 Kb windows, large and small genomic regions showing selective sweep

286   signals were identified, respectively. In total, 227 (cultivated *vs.* wild), 381 (*Ind vs.* wild), 333

287   (*Aus vs.* wild), 296 (*Aro vs.* wild), 256 (*TrJ vs.* wild) and 269 (*TeJ vs.* wild) identified regions

288   showed significantly smaller *Tajima'* D values compared to other regions (**Figure 6A**).

289   Subsequently, genes located in the selective sweep regions and their corresponding GSEA

290   (Gene Set Enrichment Analysis) terms were further identified for each subpopulation, and

291   ~50% of them were specific to each subpopulation whilst only 27 GSEA terms co-exist in the

292   five cultivated rice subpopulations (**Figure 6B**). Finally, a total of 1,180 SNPs occurred

293   within the genes in the selective sweep regions were selected to generate the fixedSNP panel.

294

295   **Subpopulation classification analysis with the fixedSNP panel**

296   To evaluate the fixedSNP panel, subpopulation classification with phylogenetic tree analysis

297   was performed using the 1,180 fixedSNPs, and the results were compared to the results

298   generated from the 156,502 tagSNPs performed on the same population of 2,556 accessions.

299   All of the accessions were assigned to the correct subpopulations with tagSNPs and the

300   phylogenetic tree showed consistent structure with the tree constructed with fixedSNPs

301   (**Figure 6C**). To further evaluate the universality of the fixedSNP panel, we performed

302   subpopulation classification on two external populations genotyped by SNP chips [11] [14].

303   One chip dataset contained 880 cultivated rice accessions genotyped by the Affymetrix 700K

304   SNP chip, while the other contained 351 cultivated accessions genotyped by the Illumina 44K

305   SNP chip. Both external chip datasets have been documented with clear subpopulation

306   classification and origins, and possess relatively high genetic diversity. Only 314 and 63

307   SNPs from the 700K and 44K chips, respectively, were found in the 1,180 fixedSNP panel.

308   For the chip dataset containing 880 accessions, 877 accessions were correctly assigned to

309   their documented subpopulations; three *TeJ* accessions (IRGC121549, IRGC121520 and

310   IRGC121535) were incorrectly assigned to the *TrJ* subpopulation (**Figure 6D**). As for chip

311   dataset containing 351 accessions, 348 were assigned to the correct subpopulation; three *TeJ*

312   accessions (NSFTV134, NSFTV204 and NSFTV283) were mistakenly assigned to *Trj* rice

313   (**Figure 6E**). Overall, 99.8% of the rice accessions examined were assigned to previously

314   documented subpopulation records using markers extracted from the fixedSNP panel,

315   indicating that the fixedSNP panel is an efficient, accurate new tool for subpopulation

316   classification.

317

318   **The barcodeSNP panel**

319   DNA fingerprinting technology using a small set of SNPs to generate a series of genotype

320   combinations, referred to as "barcodes," has become an economical means to protect

321   commercialized varieties. Thus, the barcodeSNP panel must be able to uniquely identify

322   these barcodes to distinguish between each of the rice varieties on the market. To ensure

323   highest uniqueness but lowest count of barcodeSNPs, we applied the MinimalMarker

324   algorithm on the fixedSNP panel to exhaustively traverse all possible genotype combinations

325   that would distinguish the 2,556 accessions (**Materials and Methods**). The MinimalMarker

326   algorithm generate three sets of minimum marker combinations, in which each set contains

327   28 SNPs. After merging the three sets, 38 barcodeSNPs were finally selected to generate the

328   panel (**Figure S2A**). In addition, up- and down-stream flanking sequences were also provided

329   for users to design primers for PCR-based KASP™ genotyping assays. The SR4R also offers

330   a web interface that allows users to identify corresponding accessions or varieties when rice

331   varieties are submitted for genotyping with any number of barcodeSNPs between 8 to 38.

332   The SR4R returns a list of the top 10 best-matched accessions/variety in the database, and

333   displays associated information including the accession/variety IDs, number of mismatched

334   bases, genomic position of the barcode, genotype heterozygosity, and documented

335   subpopulation and origin. Among the top 10 hits, if multiple best-matched varieties with

336   100% identity are returned using a certain number of barcodes, the users may genotype

337   additional barcodeSNPs until a unique best matched variety is identified. It is worth noting

338   that because the SR4R does not have a complete list of the barcodes for all commercial rice

339   varieties in the database, the 38 barcodeSNPs is considered as an initial panel for users to test

340   the best combinations with the most optimal sensitivity and specificity using flexible numbers

341   of markers.

342

343   **Machine learning analysis with the barcodeSNP panel**

344   If a new variety genotyped with barcodeSNPs is not found in the database, SR4R will

345   perform subpopulation classification. The traditional method of subpopulation classification

346   first integrates the genotype of the submitted variety with the genotypes of all the varieties in

347   the database, then performs phylogenetic analyses to determine the best assigned

348   subpopulation. This procedure is tedious and computationally inefficient since the database

349   contains hundreds of thousands of accessions. To simplify the procedure so that it may be

350   implemented through a web interface, we adopted an alternative method that utilizes machine

351   learning-based subpopulation classification models with the 38 barcodeSNPs as features. We

352   used all of the 2,556 rice accessions to evaluated seven commonly used machine learning

353   algorithms to perform subpopulation classification including decision tree, k-nearest

354   neighbouring, naïve Bayesian, artificial neural network, random forest, multinomial logistic

355   regression and one-*vs*-rest logistic regression algorithms, followed by ten-fold cross
356   validation assessment (**Materials and Methods**). A series of assessments of the classification
357   precision in the five cultivated rice subpopulations indicated that, out of the seven methods
358   the best one is the multinomial logistic regression model, whose AUC (Area under curve)
359   values were all ≥ 0.99 for all subpopulations (**Figure S3B-F**). Additional methods are one-vs-
360   rest logistic regression and the random forest model; where results from each yielded similar
361   classification precision to the multinomial logistic regression model. Then, we used an
362   independent datasets containing 880 rice accessions profiled by 770 Kb rice SNP chip for
363   independent validation. The multinomial logistic regression model was trained by the 2,556
364   rice accessions, and then predict the subpopulation classifications on the 880 samples. The
365   AUC values were all ≥ 0.99 for all subpopulations in this independent datasets, indicating
366   robustness of the model. Moreover, compared the original label and the predicted label with
367   the max probability for each sample, the true positive rate (TPR) and false positive rete (FPR)
368   are also reasonable (**Figure S4**). The pre-trained classification models with the seven
369   machine learning algorithms have been implemented on the SR4R server provided as a web
370   tool for users to perform subpopulation classification when the genotype information of the
371   38 barcodeSNPs in submitted.

372

The barcodeInDel panel

374   InDel (Insertion and Deletion) is another form of genomic variations (usually less than 50 bp
375   in length) that can be used as molecular markers for a variety of population analysis. From
376   the 5,152 rice accessions, a total of 4,217,174 raw InDel variations were identified using the
377   IC4R variation calling pipeline [2]. After filtering low-quality InDels, 109,898 high-
378   confidence InDels were retained with missing rate less than 0.01 and MAF ≥ 0.05 within
379   2,556 rice accessions. Among the 109,898 high-confidence InDels, we further identified 62
380   subpopulation-specific InDels which can be used as barcodeInDels to differentiate the six rice
381   subpopulation *TeJ*, *TrJ*, *Aro*, *Aus*, *Ind* and *Oru*, and the six subgroups of *indica* rice *S1-S6*
382   (**Table S3**). The 109,898 high-confidence InDels can be download from SR4R for users'
383   customized analysis.

384

Web interface of SR4R database

386   Using unified bioinformatics pipelines, the genotype data of the 18 million raw SNPs

387   identified from 5,152 rice accessions were processed to construct four reference panels of

388   SNPs for different utilizations. Because genotype data processing is a complicated and

389   computationally intensive procedure, the four SNP panels are readily usable for a variety of

390   analyses simplify task for rice researchers. For better sharing of SNPs and improvement of

391   the rice variation map utility, we developed the SnpReady for Rice (SR4R) database.

392   Through the SR4R web interface, users may directly browse the four panels and retrieve

393   detailed information related to the 2,097,405 hapmapSNPs, 125,502 tagSNPs, 1,180

394   fixedSNPs, 38 barcodeSNPs. In addition, the protein-coding genes exhibiting strong selection

395   signatures, associated with the 1,180 fixedSNPs were also included in the SR4R database

396   with detailed functional annotations (**Figure 7A**). When users retrieve a SNP such as the first

397   SNP "OSA01S00001362", the genomic location and the adjacent gene or the gene containing

398   the queried SNP are displayed. Users may also retrieve a visualized allele frequency map in

399   the six major subpopulations, and the six subgroups of *indica* rice (**Figure 7B**).

400   The users may also download the four panels of SNPs along with the original genotype

401   files for local analysis *via* http://sr4r.ic4r.org/download. In addition, the "Tools" module

402   presents 18 handy scripts and pipelines that users may install on their local computers for a

403   variety of analysis, including basic genotype processing, population diversity analysis, rice

404   variety identification and subpopulation classification. For example, assuming one user may

405   want to perform a genotype imputation of 44K SNP rice Chip, she or he may first download

406   the file "hapmapSNPs-genotype.tar.gz (892 MB)" containing the genotypes of the 2,097,405

407   hapmapSNPs in 2,516 rice accessions. Then, the user may use the pipeline and scripts

408   demonstrated in **Figure 7C** to perform imputation on a local server. SR4R also offers two

409   modules of online analysis. The first module is to use a machine learning-based method to

410   assign the subpopulation type based on the user-submitted genotype file including no more

411   than 20 samples. The model will return the probability of the type of subpopulation assigned

412   to each sample (**Figure 7D**). The second module is to perform DNA fingerprint analysis.

413   When the user submits a genotype file containing no more than 20 samples, the model will

414   search the accession database, and return the top three matches of existing varieties with the

415   number of mismatched nucleotide and heterozygosity rate displayed (**Figure 7E**). The

416   programs and scripts for these two modules along with demo input and output files are also

417   available to download for local analysis of genotypes with large sample numbers.

418

Conclusions

The IC4R Rice Variation Database collects over 18 million raw SNPs identified from resequencing of 5,152 accessions. To meet the different demands for the rice research community and breeding industry, we further generated four panels of 2,097,405 hapmapSNPs, 156,502 tagSNPs, 1,180 fixedSNPs and 38 barcodeSNPs with standard processing pipelines and uniform analytical parameters (**Table S2**). The four panels of SNPs can be either accessed online or downloaded for local use from the daughter database of RVD – SnpReady for rice (SR4R). The hapmapSNP panel contains 2 million non-missing genotypes of 2,556 accessions offers a reference HapMap for genotype imputation and high-resolution GWAS analysis. The non-redundant 150K tagSNP panel is an ideal magnitude for population genetics and evolutionary analysis for research, as well as an ideal marker pool for genomic selection-assisted breeding in rice. For a breeding population with about 500 F_1 hybrids, 1,500 to 15,000 markers selected from the tagSNP panel can be used to build a GS model, reaching a satisfactory genotype-to-phenotype prediction accuracy. The fixedSNP panel with high effectiveness and stability can be regarded as a marker pool for various molecular breeding practice suitable for low-budget, flexible genotyping platform, in terms of subpopulation classification, seed purity analysis and genetic background analysis. The 38 barcodeSNPs selected by MinimalMarker algorithm is an initial marker set for generating DNA fingerprints for commercial rice varieties. Along with the barcodeSNP panel, two web-based tools, one for variety identification and another for subpopulation classification, are offered in SR4R. In addition, the SR4R database also offers a series of standard pipelines used to construct the four sets of SNPs, and local handy tools to perform rice varieties classification, barcode development, and other types of genetic and breeding research. With the incremental accumulation of population genotype data in BIGD center, these bioinformatics tools can be applied to other animal or plant species such as corn, wheat, soybeans, for a centralized reference HapMap and SNP panel databases for plants.

Materials and Methods

Construction of hapmapSNP and tagSNP panels

The raw 18 million SNPs with genotype information of 5,152 rice accessions were obtained from the IC4R rice variation database (http://variation.ic4r.org). Accession filtration, SNP

450 filtration and basic statistics of homozygous SNPs and accession heterozygosity were
451 performed using in-house scripts. Genotype imputation of missing sites and phasing were
452 performed using Beagle [8]. A SNP site with missing genotype was removed if an inferred
453 genotype with a posterior probability was smaller than 0.5. Genomic annotation of
454 hapmapSNPs was performed using ANNOVAR (version 20160201) against the rice
455 International Rice Genome Sequencing Project (IRGSP) gene annotation. Using the reported
456 LD length of rice ranging from 40 to 500 Kb, an LD-based SNP pruning method was used to
457 construct the tagSNPs category using PLINK with *–indep* command [15] [16]. The PLINK
458 parameters were selected based on the variance inflation factor (VIF), which recursively
459 removed SNPs within a sliding window of 50 SNPs and a step size of 5 SNPs to shift the
460 window.

461

462 **Tools for subpopulation structure analysis**

463 The tagSNPs for 2,556 rice accessions were concatenated as input sequences for constructing
464 the phylogenetic tree using the neighbour joining algorithm implemented in MegaCC with
465 pairwise gap deletion and 100 bootstrap replications [17]. The output tree file for all 2,556
466 rice accessions and the subtree file of *indica* rice accessions were visualized in MEGA7 [18].
467 Principal component analysis of the 2,556 rice accessions was done by flashPCA [9].
468 Population admixture structure analysis was done by fastSTRUCTURE using the variational
469 bayesian framework, and *k=2* to *k=8* were set to infer the admixture of ancestors for the
470 accessions.

471

472 **Tools for genetic diversity analysis**

473 Genetic diversity related analyses were mostly done using PLINK [16]. Genome-wide
474 pairwise IBS calculations were performed between each pair of accessions within the same
475 subpopulation in order to deduce the genetic affinity, and an IBS pairwise distance matrix
476 was generated for each subpopulation. The ROH analysis for each subpopulation used a
477 sliding window method to scan each accession's genotype for a given population at each
478 marker position to detect homozygous segments. The parameters and thresholds applied to
479 define ROH were set as follows: a minimum ROH length of 200 Kbp and a minimum number
480 of 1,000 consecutive SNPs included in an ROH. Correlation coefficient (r^2) of SNPs was

481 calculated to measure LD level for each subpopulation. The average r^2 value was calculated

482 for each length of distance from 0 to 500 Kbp, followed by drawing LD decay figures using

483 an R script for each subpopulation. Population diversity of rice varieties was measured by

484 two indexes: $\theta\pi$ and *Fst*. Nucleotide diversity $\theta\pi$ was used as a measurement of the degree of

485 genotype variability within each subpopulation, while subpopulation differentiations were

486 evaluated by the fixation index *Fst* for each of the cultivated subpopulations against the wild

487 rice population and for the cultivated subpopulations compared to each other. Values of $\theta\pi$

488 and *Fst* were calculated using sites mode implemented in VCFtools [19].

489

490 **Tools for genomic selection analysis**

491 Genotype and phenotype datasets of the 44k rice chip were downloaded from the Rice

492 Diversity Website (http://www.ricediversity.org/). Genotype imputation and phasing were

493 then performed using Beagle (version 3.3.2), and the site was filtered if an inferred genotype

494 with a posterior probability was smaller than 0.5. Genomic selection analysis was performed

495 using RR-BLUP mixed model implemented in R package rrBLUP [13] for nine well-

496 measured traits (flowering time, panicle fertility, seed width, seed volume, seed surface area,

497 plant height, flag leaf length, flag leaf width, and florets per panicle) with five different

498 feature combinations. The prediction accuracy under each feature combination was evaluated

499 by five-fold cross-validation and Pearson correlation coefficient. An example of the process

500 is as follows: the original samples were randomly partitioned into five subsets; of the five

501 subsets, a single subset was retained as the validation data, and the remaining four subsets

502 were used as training data. This process was repeated five times, with each of the ten subsets

503 used exactly once as the validation data. The Pearson correlation coefficients of the predicted

504 breeding values and the real phenotype values were calculated for each fold.

505

506 **Construction of the fixedSNP panel**

507 $\theta\pi$ and *Tajima'* D values were calculated for six rice subpopulations (*TeJ, TrJ, aro, aus, ind,*

508 *Oru*) with a sliding-window fashion across the genome using in-house scripts. *Fst* values

509 were calculated for the five cultivated subpopulations against the wild *Oru* subpopulation, as

510 well as for the five cultivated subpopulations against each other. For each pairwise

511 comparison, the intersection of the top 5% windowed $\theta\pi$ ratios (wild subpopulation *vs.*

512 cultivated subpopulation), and the top 5% windowed *Fst* values correspondingly were
513 selected as strong selective sweep signals. Window sizes of both 100 Kbp and 10 Kbp were
514 used to detect large or small selective sweep regions, followed by merging the results as the
515 candidate selective sweep regions for each subpopulation. *Tajima'* D distribution was also
516 drawn for the candidate selective sweep regions against the whole genomes for each pairwise
517 comparison. Genes located within the candidate selective sweep regions were extracted for
518 each comparison, and Gene Set Enrichment Analysis (GSEA) was performed for each gene
519 listed by using PlantGSEA web tools [20]. Genic SNPs located in the candidate selective
520 sweep regions identified from the above-mentioned pairwise comparisons were merged as
521 fixedSNPs.

522

Construction of the barcodeSNP panel

524 The 1,180 fixedSNPs were used as the initial marker set to select the minimal number of
525 barcodeSNPs that can maximally distinguish the 2,556 rice accessions using a heuristic mode
526 implemented in MinimalMarker [7]. Three minimal sets each containing 28 SNPs were
527 generated, and after merging the three sets, 38 unique SNPs were selected as barcodeSNPs
528 for generating DNA fingerprints for each accessions.

529 To identify commercialized rice varieties using the combination of 38 barcodeSNPs, seven
530 machine learning-based methods were used: decision tree, k-nearest neighboring, naïve
531 Bayesian, artificial neural network, random forest, multinomial logistic regression, and one-
532 vs-rest logistic regression algorithms in the Python sklearn library (https://scikit-
533 learn.org/stable/). The precision of each model was assessed using ten-fold cross-validation
534 method. Specifically, the original sample set was randomly partitioned into ten subsets in
535 which nine subsets were used for training model and the remaining subset was used as the
536 testing model; this procedure was repeated ten times and an average prediction accuracy was
537 computed from the overall performance of the tested models. Five one-hot codes (10000,
538 01000, 00100, 00010, 00001) to label the five subpopulations for classification using
539 machine learning models. Then, the predicted label with the max probability was compared
540 with the original label for each sample. If the predicted label is identical with the original
541 label, the prediction result was regarded as correct. Then, the ratios of positive and negative
542 rate were computed to plot ROC curves and compute AUC values.

543

Construction of the barcodeInDel panel

Raw InDels were identified using the IC4R variation calling pipeline from the origin 5,152 rice accessions [2]. Then, the InDels from the 2,556 rice accessions with high sequencing coverage (depth $\geq$ 5) presented in SR4R database were extracted using customized Python scripts, followed by using VCFtools [19] to filter InDels to generate a high-confidence InDel dataset, with parameters of missing rate less than 0.01 and MAF $\geq$ 0.05. Finally, using customized Python scripts, InDels which have the same sequence type within each subpopulation were retained to generate the subpopulation-specific barcodeInDel panel.

552

Data availability

All the data is freely available and downloadable at http://sr4r.ic4r.org/.

555

Authors' contributions

XFW, SHS and ZZ conceived the project; JY and CL collected the samples; JY conducted the data analysis; DZ developed the database; JY, XFW, SHS and ZZ wrote the manuscript.

559

Competing interests

The authors declare no competing interests.

562

Acknowledgments

569

References

571 [1] Li Z, Fu BY, Gao YM, Wang WS, Xu JL, Zhang F, et al. The 3,000 rice genomes
572   project. GigaScience 2014;3.

573 [2] Zhang Z, Hu S, He H, Zhang H, Chen F, Zhao W, et al. Information Commons for
574   Rice (IC4R). Nucleic Acids Research 2016;44:D1172–80.

575 [3] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second
576   generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61.

577 [4] Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of Linkage Disequilibrium in
578   Plants. Annual Review of Plant Biology 2003;54:357–74.

579 [5] Nielsen R. Molecular Signatures of Natural Selection. Annual Review of Genetics
580   2005;39:197–218.

581 [6] Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, et al. Resequencing of 31 wild
582   and cultivated soybean genomes identifies patterns of genetic diversity and selection.
583   Nature Genetics 2010;42:1053–9.

584 [7] Fujii H, Ogata T, Shimada T, Endo T, Iketani H, Shimizu T, et al. Minimal marker: An
585   algorithm and computer program for the identification of minimal sets of
586   discriminating dna markers for efficient variety identification. Journal of
587   Bioinformatics and Computational Biology 2013;11:1250022.

588 [8] Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data
589   inference for whole-genome association studies by use of localized haplotype
590   clustering. American Journal of Human Genetics 2007;81:1084–97.

591 [9] Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide
592   data. PLoS ONE 2014;9.

593 [10] Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in
594   3,010 diverse accessions of Asian cultivated rice. Nature 2018;557:43–9.

595   [11]   McCouch SR, Wright MH, Tung CW, Maron LG, McNally KL, Fitzgerald M, et al.
596   Open access resources for genome-wide association mapping in rice. Nature
597   Communications 2016;7.

598   [12]   Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, et al. Genomic
599   Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic
600   Architecture, Training Population Composition, Marker Number and Statistical Model
601   on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS
602   Genetics 2015;11:1–25.

603   [13]   Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R
604   Package rrBLUP. The Plant Genome Journal 2011;4:250.

605   [14]   Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, et al. Genome-wide
606   association mapping reveals a rich genetic architecture of complex traits in Oryza
607   sativa. Nature Communications 2011;2.

608   [15]   Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. The
609   extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics 2007;177:2223–32.

610   [16]   Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-
611   generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience
612   2015;4:7.

613   [17]   Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of
614   molecular evolutionary genetics analysis program for automated and iterative data
615   analysis. Bioinformatics 2012;28:2685–6.

616   [18]   Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis
617   Version 7.0 for Bigger Datasets. Molecular Biology and Evolution 2016;33:1870–4..

618   [19]   Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The
619   variant call format and VCFtools. Bioinformatics 2011;27:2156–8.

620   [20]   Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant
621   community. Nucleic Acids Research 2013;41:W98-103.

622

Figure legends

Figure 1 An overview of the four SNP panels of the SR4R database

The flow chart describes procedures on how the four SNP panels were generated.

Figure 2 Basic statistics of the rice hapmapSNPs after four steps of genotype processing

After four steps of genotype processing, a series of basic statistical analyses was performed at each step to exhibit the characteristics of the SNPs, including: **A**. statistics of individual missing rate, **B**. statistics of individual heterozygote rate, **C**. statistics of minor allele frequency, **D**. statistics of site missing rate, and **E**. statistics of site heterozygote rate. **F.** After ARNOVAR analysis to annotated the hapmapSNPs, distribution of the hapmapSNPs in different genomic regions were analyzed.

Figure 3 Population structure analysis of the 2,556 rice accessions using tagSNPs

To test whether the 150K tagSNPs can generate the population structures consistent with previous reports, we performed a series of population structure analyses to generate: **A**. the phylogenetic tree, **B.** the PCA map, **C.** the admixture structure of 2,556 rice accessions, **D.** the phylogenetic tree of the six subgroups of *indica* rice. The tagSNPs effectively and accurately classified the 2,556 rice accessions to corresponding populations.

Figure 4 Genetic diversity analysis of rice accessions using tagSNPs

The 150K tagSNPs were used in a series of population genetic analysis to show the effectiveness of tagSNPs including: **A**. statistics of homozygous SNPs, **B.** statistics of individual heterozygosity, **C**. pairwise IBS values distribution, **D.** statistics of ROH regions, **E**. LD decay analysis, in the five major rice subpopulations. **F**. Genetic diversity ($\theta\pi$) and population differentiation (*Fst*) between cultivated and wild subpopulations. **G**. Population differentiation (*Fst*) of cultivated subpopulations.

Figure 5 Genomic selection-based phenotype prediction using tagSNPs

Nine agronomical phenotypes were predicted based on rrBLUP models to evaluate the effectiveness of tagSNPs. Five sets of SNPs with equal amounts were compared, including Set-1: the original 29,434 SNPs on the 44K chip; Set-2: the 1,090 SNPs overlapped between the 156,502 tagSNPs and 29,434 SNPs; Set-3: the 1,090 SNPs randomly selected from the 29,434 SNPs; Set-4: the 1,090 SNPs evenly distributed in the genome (350 Kb per SNP)

656 selected from the 29,434 SNPs; Set-5: 1,090 SNPs localized within a genomic region from

657 the 29,434 SNPs.

658

659 **Figure 6 Screening and validation of FixedSNPs**

660 **A**. Distribution of $\theta\pi$ ratios (wild vs cultivar) and corresponding *Fst* values, which are

661 calculated in 100kb windows. Data points located to the right of the vertical dashed line and

662 to the top of the horizontal dashed line are potential strong selective sweep signals (Red

663 points, corresponding to the 5% right tails of the empirical $\theta\pi$ ratio and *Fst* values distribution,

664 respectively). Distribution of Tajima's D values for the potential selective sweep signals and

665 whole genomes are shown within the plot. Other comparisons for the screening of subgroup

666 specific selective sweep signals were not shown here, but demonstrate similar trends. **B**.

667 Common and specific selective signals among cultivar subgroups (Number of genes or GSEA

668 terms are shown out and in the brackets, respectively). **C**. Phylogenetic tree of 2,538 rice

669 cultivars in fixedSNPs data set. **D**. Phylogenetic tree of 880 rice cultivars in 700K chip data

670 set. **E**. Phylogenetic tree of 351 rice cultivars in 44K chip data set.

671

672 **Figure 7 Representative functional modules in SR4R database**

673 **A.** Genes exhibiting significant selection signatures in the corresponding subpopulations are

674 listed in the "Selected Genes" module in the Browser. **B.** Allele frequencies in different

675 subpopulations of the first hapmapSNP (SNPID: OSA01S00001362, associated gene:

676 Os01g0100100, position: chr01-1362, allele: Alt-A, Ref-G). **C**. One example of the script and

677 pipeline for population diversity analysis. **D**. The online analysis module of subpopulation

678 classification using machine learning algorithms. **E**. The online analysis module of rice

679 variety identification using the 38 barcodeSNPs.

680

681 **Supplementary material**

682 **Figure S1 Population structure of 1,655 varieties in *indica***

683 **A**. PCA classification for 1,655 varieties in *indica* subgroup. **B**. Structure analysis for 1,655

684 varieties in *indica* subgroup.

685

686 **Figure S2 T-test for Pearson correlations of the selected 1,090 tagSNPs set and other**

687 **four SNP sets**

688 Set-1: the original 29,434 SNPs on the 44K chip; Set-2: the 1,090 SNPs overlapped between

689 the 156,502 tagSNPs and 29,434 SNPs; Set-3: the 1,090 SNPs randomly selected from the

690 29,434 SNPs; Set-4: the 1,090 SNPs evenly distributed in the genome (350 Kb per SNP)

691 selected from the 29,434 SNPs; Set-5: 1,090 SNPs localized within a genomic region from

692 the 29,434 SNPs. Different colors present different P values. * P value < 0.05; ** P value <

693 0.01.

694

695 **Figure S3 BarcodeSNPs and machine learning models for the classification of rice**

696 **varieties**

697 **A**. Heat-map of BarcodeSNPs of 2,538 rice cultivars (Red: A, Yellow: T, Blue: G, Green: C).

698 **B**. Decision model. **C**. KNN model. **D**. Multinomial logistic regression model. **E**. Naive

699 Bayesian model **F**. Neural network model. **G**. Random forest model. **H**. One-vs-rest logistic

700 regression model. AUC curves were drawn using the mean values of ten cross validations for

701 B-H.

702

703 **Figure S4 Independent validation of the machine learning model.**

704 **A**. ROC curve for the 770 Kb rice SNP chip dataset using the pre-build multinomial logistic

705 regression model. **B.** The true positive rate (TPR) and false positive rate (FPR) statistics for

706 each subpopulation of the 770 Kb rice SNP chip dataset.

707

708

709 **Table S1   Summary of 2,556 rice accessions with subpopulation classification and**

710 **origins**

711

712 **Table S2 Summary of SNPs annotation for SR4R database**

713

714 **Table S3 The barcodeInDel panel in SR4R database**

715

716

717   **Figure 1**
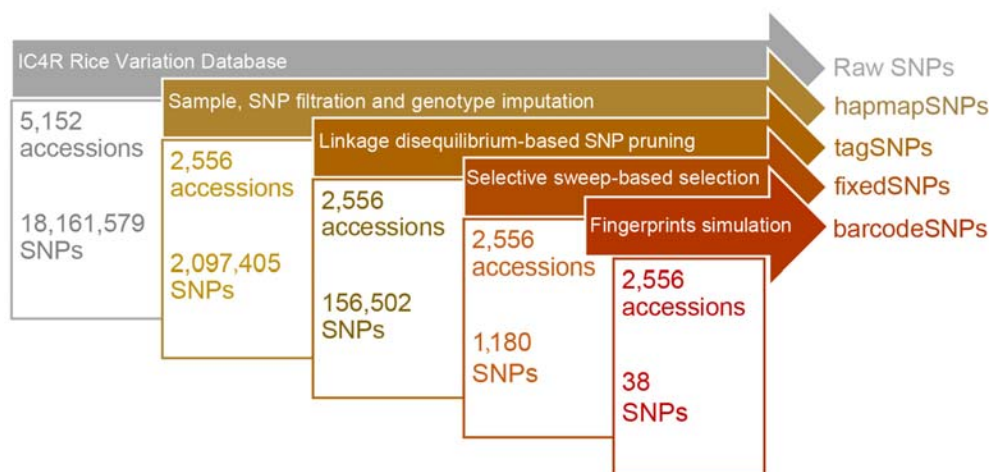
718
719
720
721
722
723

724   **Figure 2**
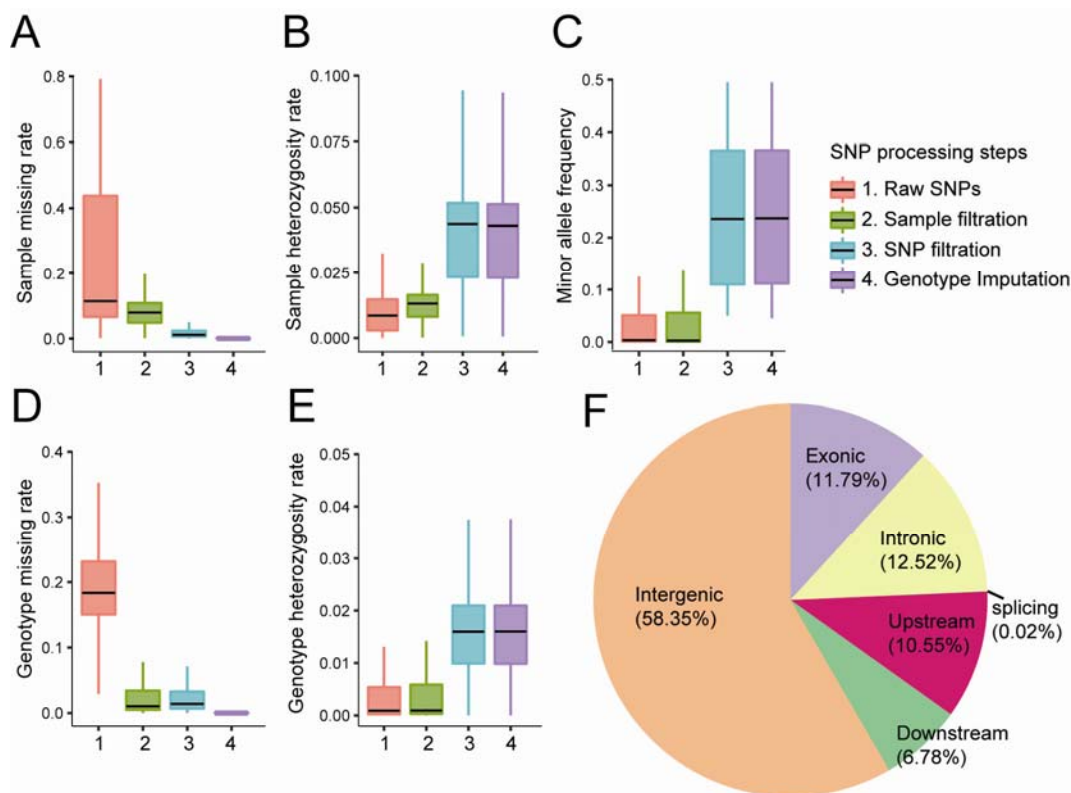
725
726

727   **Figure 3**
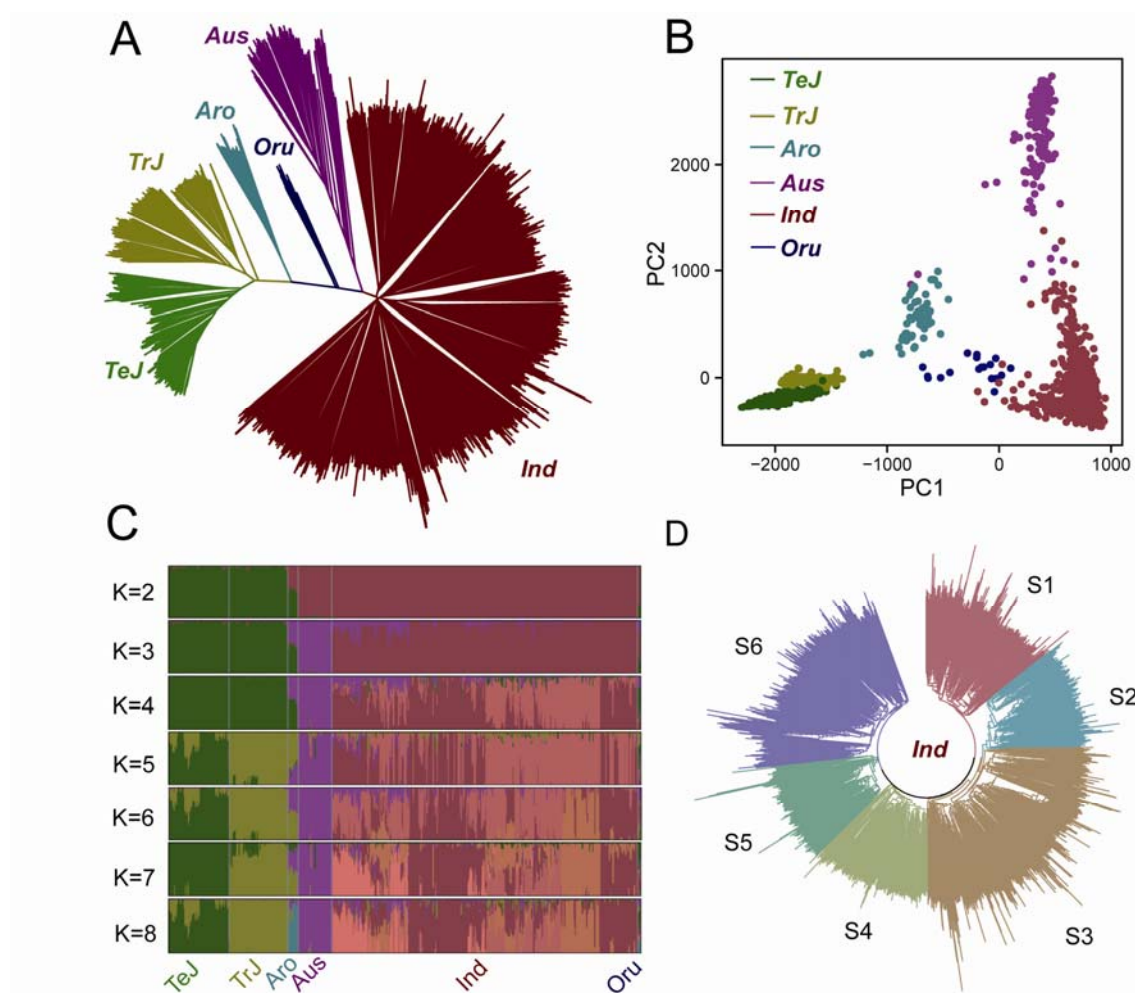
728

729

730

731

732

733

734

735

736

737

738

739

740

741

742 **Figure 4**
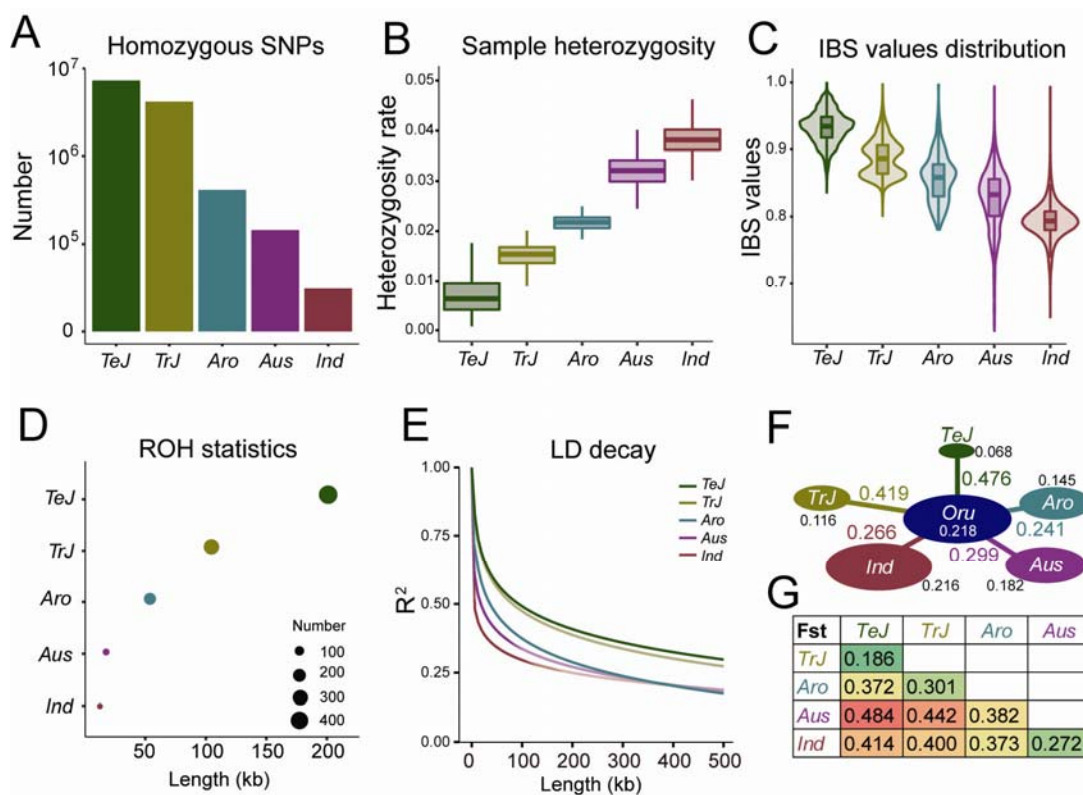
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759

760 **Figure 5**
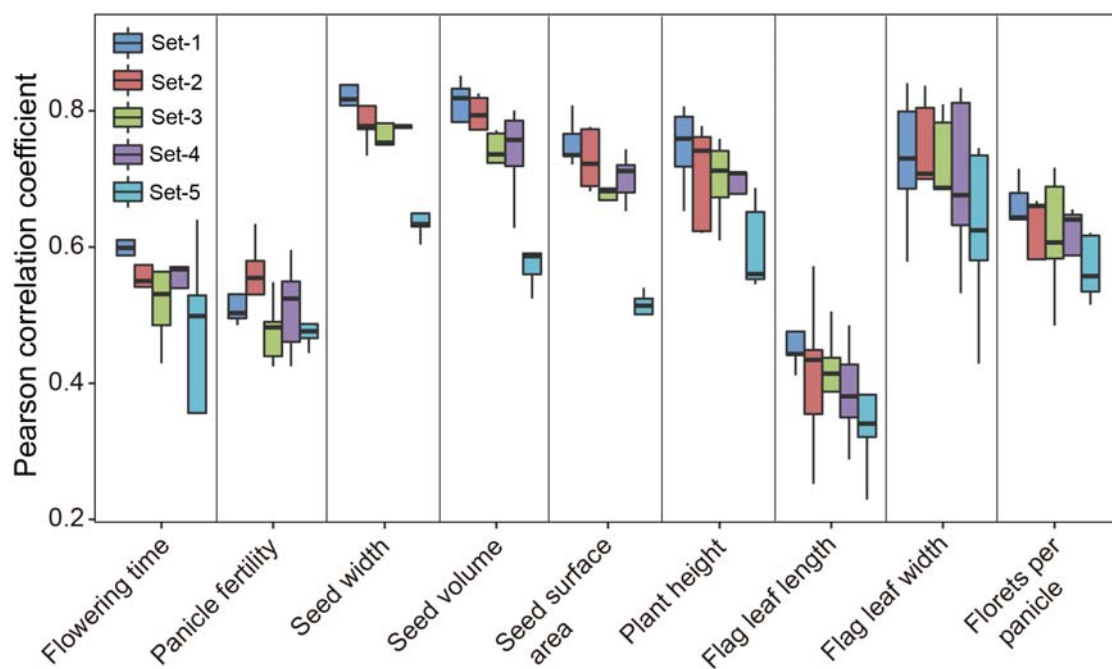
781   **Figure 6**
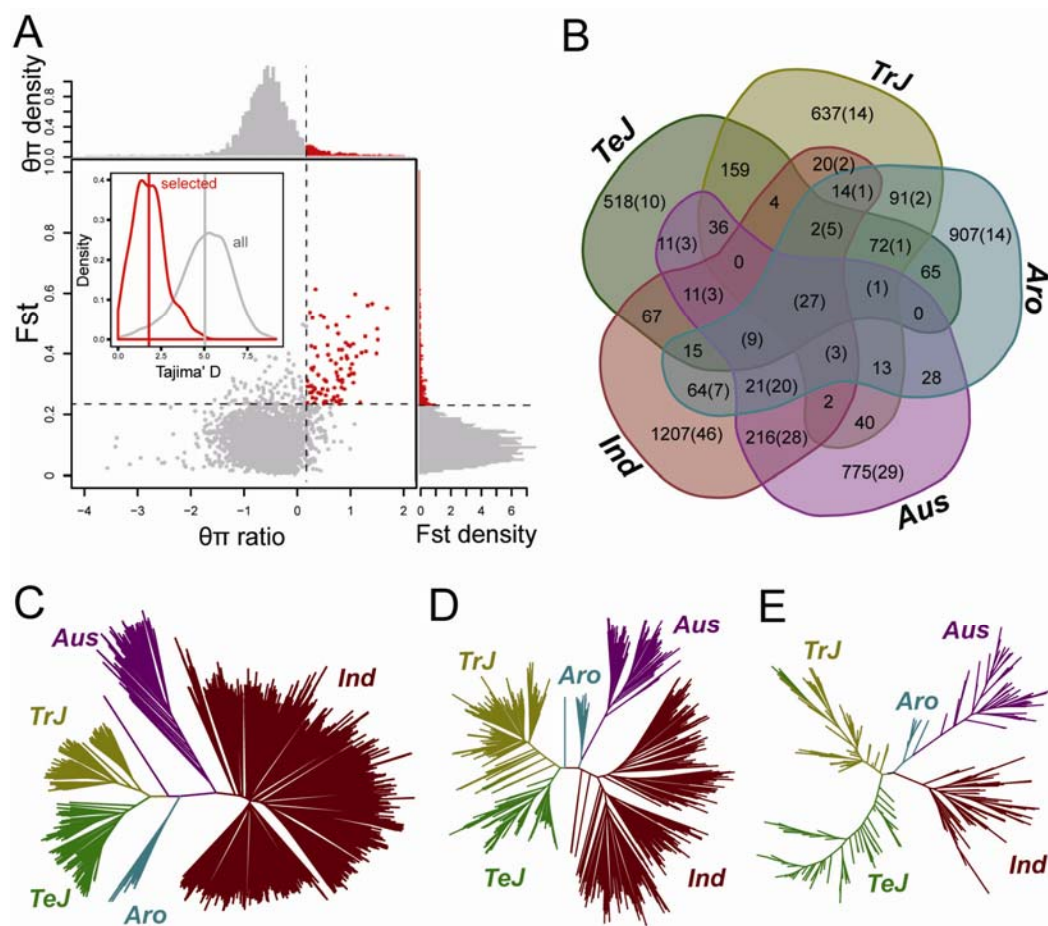
782
783
784
785
786
787
788
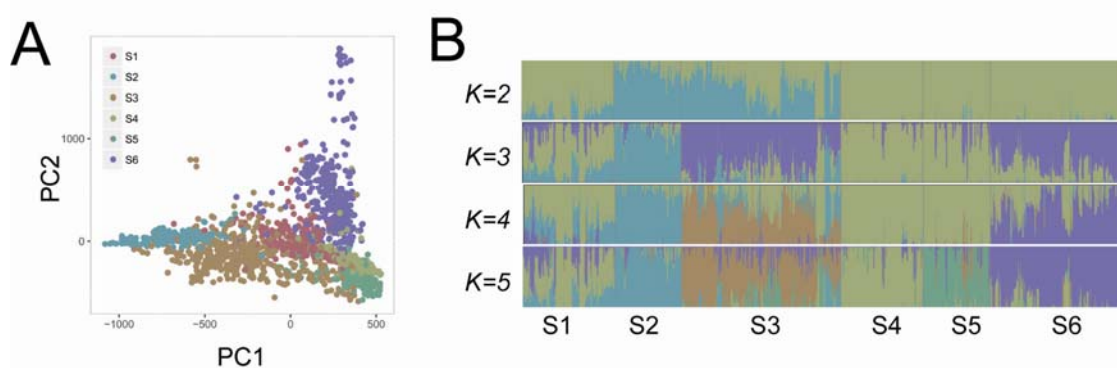789
790
791
792
793
794
795
796

797   **Figure 7**

798
799
800
801
802
803

804 **Figure S1**

805

806

807

808

809

810

811 **Figure S2**

	Flowering time	Panicle fertility	Seed width	Seed volume	Seed surface area	Plant height	Flag leaf length	Flag leaf width	Florets per panicle
Set-1 vs Set-2	0.2279	0.7783	0.0265 *	0.0243 *	0.0327 *	0.0346 *	0.1081	0.1834	0.0621
Set-2 vs Set-3	0.2273	0.0595	0.0212 *	0.0042 **	0.0369 *	0.3687	0.4189	0.0090 **	0.6542
Set-2 vs Set-4	0.3961	0.0485 *	0.0514	0.0008 **	0.0562	0.2971	0.2035	0.1793	0.0362 *
Set-2 vs Set-5	0.0127 *	0.0728	0.0015 **	0.0003 **	0.0002 **	0.0140 *	0.1779	0.0004 **	0.1819

P value significance

- 0.0001 to 0.001
- 0.001 to 0.01
- 0.01 to 0.05
- 0.05 to 0.1
- 0.1 to 0.2
- 0.2 to 0.3
- 0.3 to 0.4
- 0.4 to 0.5
- 0.5 to 1.0

812

813

814 **Figure S3**

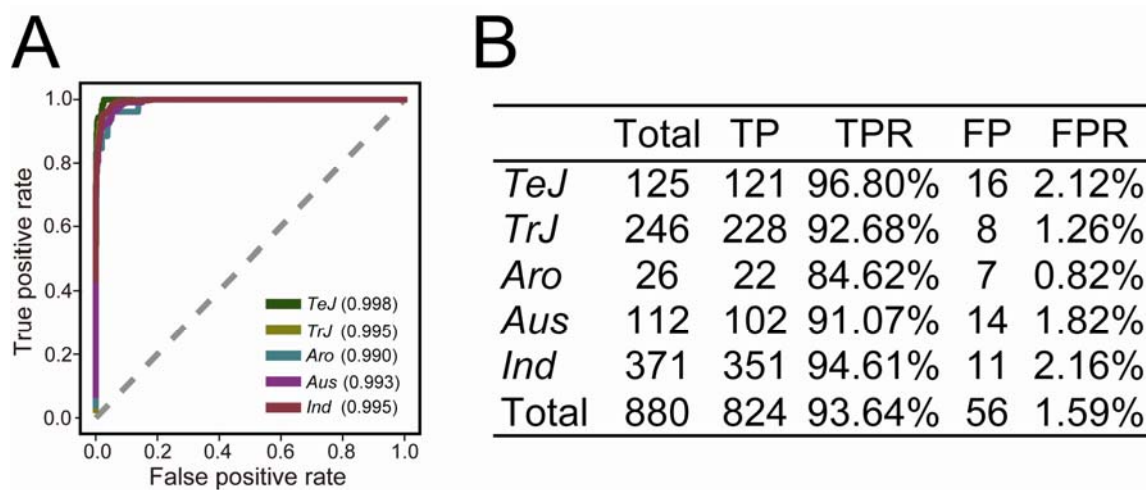815
816
817

818 **Figure S4**

	Total	TP	TPR	FP	FPR
TeJ	125	121	96.80%	16	2.12%
TrJ	246	228	92.68%	8	1.26%
Aro	26	22	84.62%	7	0.82%
Aus	112	102	91.07%	14	1.82%
Ind	371	351	94.61%	11	2.16%
Total	880	824	93.64%	56	1.59%

819