Running title (optional): **Choosing the optimal population for GWAS**

Core ideas (3-5 impact statements, 85 char max for each)

- Genome-wide association studies with mixture populations are expected to improve the detection power of novel genes due to the increase of the sample size although the influence of population structure is a concern.
- When a quantitative trait nucleotide (QTN) is polymorphic in a target population, a combination of the target population and a population with higher diversity than the target population improves the detection power of the QTN.
- We found that the fixation index (F_{ST}) and the expected heterozygosity (H_e) were strongly related to the detection power of QTNs.
- Germplasm collections which have been already sequenced/genotyped are useful for improving the detection power of GWAS without any addition of sequence costs by using a subset of them with a target population.

Choosing the optimal population for a genome-wide association study: a simulation using whole-genome sequences from rice

Kosuke Hamazaki, Hiromi Kajiya-Kanegae, Masanori Yamasaki, Kaworu Ebana, Shiori Yabe, Hiroshi Nakagawa and Hiroyoshi Iwata*

Affiliations:

K. Hamazaki, H. Kajiya-Kanegae and H. Iwata, Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan; H. Kajiya-Kanegae, current address: Research Center for Agricultural Information Technology, National Agriculture and Food Research Organization, 3-1-1 Kannondai, Tsukuba, Ibaraki 305-8517, Japan; M. Yamasaki, Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, 1348 Uzurano, Kasai, Hyogo 675-2103, Japan; K. Ebana, Genetic Resources Center, National Agriculture and Food Research Organization, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan; S. Yabe, Institute of Crop Science, National Agriculture and Food Research Organization, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8518, Japan; H. Nakagawa, Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, 3-1-3 Kannondai, Tsukuba, Ibaraki 305-8604, Japan. *Corresponding author (aiwata@mail.ecc.u-tokyo.ac.jp)

Abbreviations:

AUC, area under the curve; CDR, correct detection rate; FDR, false discovery rate; FN, false negative FP, false positive; F_{ST}, the fixation index; GWAS, genome-wide association study; H, high; HM, higher-middle; H_e, the expected heterozygosity; L, low; LD, linkage disequilibrium; LM, lower-middle; M, middle MAF, minor allele frequency QTL, quantitative trait loci; QTN, quantitative trait nucleotide; ROC, receiver operating characteristic; SNP, single nucleotide polymorphism; TN, true negative; TP, true positive.

ABSTRACT

A genome-wide association study (GWAS) needs to have a suitable population. The factors that affect a GWAS, e.g. population structure, sample size, and sequence analysis and field testing costs, need to be considered. Mixture populations containing subpopulations of different genetic backgrounds may be suitable populations. We conducted simulation experiments to see if a population with high genetic diversity, e.g., a diversity panel, should be added to a target population, especially when the target population harbors small genetic diversity. The target population was 112 accessions of *Oryza sativa* subsp. *japonica*, mainly developed in Japan. We combined the target population with three populations that had higher genetic diversities. These were 100 *indica* accessions, 100 *japonica* accessions, and 100 accessions with various genetic backgrounds. The results showed that the GWAS power with a mixture population was generally higher than with a separate population. Also, the GWAS optimal population varied depending on the fixation index F_{ST} of the quantitative trait nucleotide (QTN) and its polymorphism of QTN in each population. When a QTN is polymorphic in a target population, a target population combined with a higher diversity population improves the

59 QTN detection power. Investigating F_{ST} and the expected heterozygosity H_e as factors
60 influencing the detection power, we showed that SNPs with high F_{ST} or low H_e are less likely to
61 be detected by GWAS with mixture populations. Sequenced/genotyped germplasm collections
62 can improve the GWAS detection power by using a subset of them with a target population.

63

64 **INTRODUCTION**

65 Recently, as genome sequencing costs have continued to decrease (Metzker, 2010), the
66 whole-genome sequences of a large number of cultivars/lines have become available for major
67 crop species, such as rice (Li et al., 2014; Wang et al., 2018). A genome-wide association study
68 (GWAS) based on whole-genome sequences can more efficiently and accurately identify genes
69 that control important agronomic traits than previous methods (Koboldt et al., 2013; Ott et al.,
70 2015; Yano et al., 2016).

71 It is important to prepare an appropriate population to be analyzed when attempting to detect
72 candidate genes using GWAS techniques. For example, to avoid potential false positives caused
73 by population stratification/structure, a GWAS population should be selected that results in low
74 stratification (Begum et al., 2015; Yano et al., 2016). However, if such a population is selected as
75 an analytical population for a GWAS, the sample size may be limited and the detection power of
76 the GWAS will decrease (Korte and Farlow, 2013). Therefore, when designing an appropriate
77 GWAS population, one should be aware of the trade-off relationship between population
78 stratification and sample size.

79 When preparing the population to be analyzed, the factors that directly affect the GWAS
80 results, such as population structure, sample size, and the sequence analysis and cultivation

testing costs, need to be considered. In recent years, the whole-genome sequences of a large number of cultivars/lines have become publicly available due to highly efficient sequencing analyses and database enrichment. The publically available whole-genome sequence data will improve GWASs and could enable researchers to avoid the costs of sequencing analyses. For example, in rice, "The 3,000 Rice Genomes Project" (Li et al., 2014; Wang et al., 2018) by the International Rice Research Institute (IRRI) is a well-known whole-genome sequence dataset that is available in the "Rice SNP-Seek Database" (Alexandrov et al., 2015; Mansueto et al., 2016; 2017). Therefore, an appropriate GWAS population could potentially utilize existing public sequence data.

A mixture population obtained by mixing subpopulations with different genetic backgrounds could also potentially be used in a GWAS. An advantage of using such a mixture population is that it should improve the detection of causal variants by increasing the sample size. Conversely, a GWAS with a mixture population may suffer from large numbers of false positives caused by the population structure. Although a mixed effect model that suppresses the influence of the population structure has been proposed (Yu et al., 2006), such a mixture population has rarely been analyzed by a GWAS.

An actual data analysis of rice using whole-genome sequences showed that the detection power of a GWAS improved when *Oryza sativa* subsp. *japonica* and *Oryza sativa* subsp. *indica* populations were combined (Misra et al., 2017). Furthermore, the identification of new rice genes using a GWAS and populations with extremely high genetic diversities has also been previously reported (Zhao et al., 2011). Conversely, it has been reported that the genetic differentiation between subpopulations in a population with high genetic diversity could cause a reduction in the power of a GWAS (Huang et al., 2012). Therefore, real data studies have been

inconsistent about whether mixture populations or populations with high genetic diversities should be used in a GWAS. However, these previous studies mostly analyzed actual data, and there have been no theoretical simulation studies that have considered the possibility of using a mixture population in a GWAS. Furthermore, no previous studies have discussed which kinds of populations should be mixed to improve the GWAS detection power or which kinds of populations are most appropriate for a GWAS. Therefore, in this study, we conducted simulation experiments to see whether adding a population with a high genetic diversity compared to a target population (e.g., adding a diversity panel to a target population) is appropriate, especially when the genetic diversity of the target population is small.

MATERIALS AND METHODS

Materials (populations used in the GWAS)

In this study, 112 accessions of *Oryza sativa* subsp. *japonica* (referred to as "A"), which were accessions that had mainly been developed in Japan, were used as a target population with low genetic diversity (Yabe et al., 2016). We used the following three populations selected from "The 3,000 Rice Genomes Project" (J. Y. Li et al., 2014), i.e., 100 accessions of *Oryza sativa* subsp. *indica* (referred to as "B"), 100 accessions of *Oryza sativa* subsp. *japonica* (referred to as "C" or temperate), and 100 accessions of *Oryza sativa* with various genetic backgrounds (referred to as "D" or diverse), as populations with higher diversities than the target population (Table S1 in Supplemental File 1). The process used to select each of the 100 accessions is described in Supplemental File 9. One accession (IRIS ID: IRIS 313-11868) was duplicated in populations B and D. Among populations B, C, and D, the B population was the most

126 differentiated from A, whereas C was the most similar to A. Population D contained subsp.
127 *indica*, subsp. *japonica*, and *aus*, and *aromatic* rice accessions, which meant that the D
128 population had the highest genetic diversity. Fig. 1 is an unrooted phylogenetic tree that shows
129 the genetic relationships among accessions belonging to populations A, B, C, and D.
130

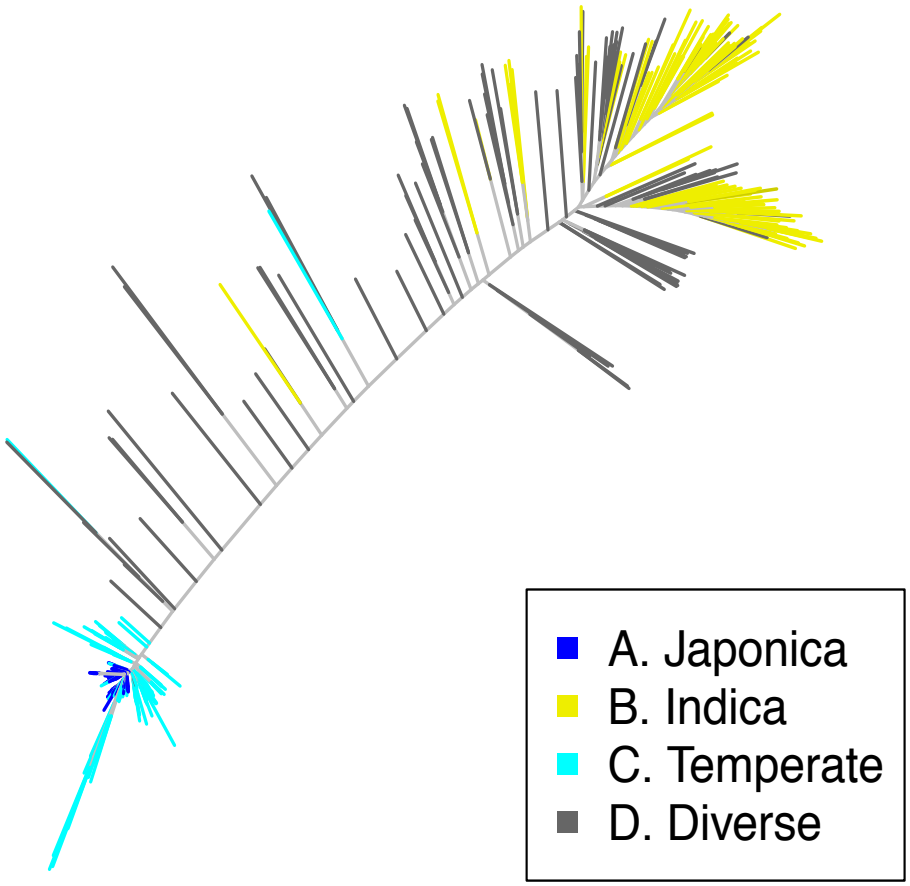


Fig. 1. **Unrooted phylogenetic tree plot for four non-mixture populations.**

Unrooted phylogenetic tree plot for the four non-mixture populations, which consisted of 112 accessions of *japonica* (A), 100 accessions of *indica* (B), 100 accessions of temperate *japonica* (C), and 100 diverse accessions (D) with neighbor-joining method.

The genetic relationships among the accessions were estimated by the neighbor-joining (NJ) method (Saitou and Nei, 1987) using the R package "ape" version 5.3 (Paradis et al., 2004). The genetic distances were estimated according to the Jukes and Cantor (1969) model. In addition to these four populations, we synthesized three populations by combining population A with populations B, C, or D. The mixture populations A + B, A + C, and A + D were named "E", "F", and "G", respectively. We compared the QTN detection power the GWAS when the seven non-mixture (A, B, C, and D) and mixture populations (E, F, and G) were used.

Genotype data

Whole genome sequencing data were available for the accessions (Jarquin et al., 2019). Details about the DNA extraction and whole genome sequencing techniques are provided in a previous report (Jarquin et al., 2019). The data sets deposited in the DDBJ Sequence Read Archive (SRA106223, ERA358140, DRA000158, DRA000307, DRA000897, DRA000927, DRA007273, DRA007256, and DRA008071) were reanalyzed. We processed the whole-genome sequence data as follows so that they could be used in the GWAS. Adapters and low-quality bases were removed from paired reads using the Trimmomatic v0.36 program (Bolger et al., 2014). The preprocessed reads were aligned using Os-Nipponbare-Reference-IRGSP-1.0 (Kawahara et al., 2013) and the bwa-0.7.12 mem algorithm with the default options (H. Li, 2012). The binary alignment map (BAM) files deposited in the Rice SNP-Seek database were also reanalyzed. Single nucleotide polymorphism (SNP) calling was based on alignments determined using the Genome Analysis Toolkit (GATK), 3.7-0-gcfedb67 (McKenna et al., 2009; Auwera et

159 al., 2014) and Picard package V2.5.0 (http://broadinstitute.github.io/picard). The mapped reads
160 were realigned using RealignerTargetCreator and indelRealigner in the GATK software. The
161 SNPs and InDels were called at the population level using the UnifiedGenotyper in GATK and
162 the -glm BOTH option. We extracted bi-allelic sites in all the accessions from the variants using
163 VCFtools version 0.1.13 (Danecek et al., 2011). Then, imputations were imputed using Beagle
164 version 4.1 (Browning and Browning, 2016). Finally, we analyzed the SNPs with minor allele
165 frequencies (MAFs) ≥ 0.05 in each population. In the analysis, the genotypes were represented as
166 -1 (homozygous of the reference allele), 1 (homozygous of the alternative allele) or 0
167 (heterozygous of the reference and alternative alleles). Out of all the whole-genome sequence
168 polymorphisms, only the SNPs on chromosome 1 were analyzed. The number of SNPs on
169 chromosome 1 in each population is shown in Table 1.

170

171 **Table 1. Number of SNPs and the diversity level of non-mixture and mixture populations.**

	Population name	Number of accessions	Number of SNPs	Diversity level†
A.	Japonica	112	72,110	263.095
B.	Indica	100	427,943	660.416
C.	Temperate japonica	100	135,665	362.649
D.	Diverse	100	647,731	798.646
E.	A + B	212	633,507	803.064
F.	A + C	212	151,675	334.606
G.	A + D	212	684,774	859.678

172 † Diversity level is the index that was used to indicate the degree of genetic diversity and is
173 described in the "Degree of genetic diversity index" section below.

Generating phenotype data

Phenotypic data were simulated using the following formula:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \mathbf{u} + \mathbf{e}, \quad \text{(Eq. 1)}$$

where $\mathbf{y}$ is the vector that represents the simulated phenotypic values for all 411 accessions; $\mathbf{X}$ is the design matrix representing the genotypes of three quantitative trait nucleotides (QTNs) with scores −1, 0, or 1; $\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \beta_3]^{\mathrm{T}}$ is the vector representing the effects of the three QTNs, $\mathbf{u}$ is the vector for polygenetic effects, and $\mathbf{e}$ is the residuals vector. Three QTN-SNPs whose MAF was equal to or larger than 0.05 in all 411 accessions (672,923 SNPs in total) were randomly selected from the SNPs on chromosome 1. The simulations were divided into five categories (low, lower-middle, middle, higher-middle, high) based on the fixation index (F_{ST}) between populations A and B for the first QTN (Fig. S1 in Supplemental File 2). We assumed that the first QTN had four times greater variance than the remaining two QTNs (referred to as "QTN1", "QTN2", and "QTN3" respectively). The remaining two QTNs were chosen randomly from SNPs where the F_{ST} between A and B were low (SNPs whose F_{ST} value was in the lower 20% category among the 672,923 SNPs). The F_{ST} for each marker was calculated according to Wright (1965) as follows:

$$F_{ST} = 1 - \frac{H_S}{H_T}, \quad \text{(Eq. 2)}$$

where H_S is the average of the expected heterozygosity based on the allele frequencies of populations A and B, and H_T is the expected heterozygosity based on the average allele frequency of populations A and B. H_S and H_T were calculated as follows:

$$H_S = \frac{N_A \cdot \{2p_A(1-p_A)\} + N_B \cdot \{2p_B(1-p_B)\}}{N_A + N_B}, \quad \text{(Eq. 3)}$$

$$H_T = 2\left(\frac{N_A p_A + N_B p_B}{N_A + N_B}\right)\left(1 - \frac{N_A p_A + N_B p_B}{N_A + N_B}\right), \quad \text{(Eq. 4)}$$

where p_A, p_B, N_A, and N_B are the allele frequencies and the sample sizes of populations A and B respectively, and $N_A = 112$ and $N_B = 100$. The F_{ST} distribution between A and B is shown in Fig. S1, which also shows the thresholds for the five F_{ST} categories.

The polygenetic effect in Eq. 5 was sampled from the multivariate normal distribution whose variance-covariance matrix was proportional to the additive numerator relationship matrix **A** and was normalized so that their variance was equal to that of the three QTN effects.

$$\mathbf{u} \sim \mathrm{MVN}(0, \mathbf{G}), \quad \text{(Eq. 5)}$$

where $\mathbf{G} = \mathbf{A}\sigma_A^2$ is the genetic covariance matrix, and the additive genetic variance σ_A^2 was automatically determined from the relationship with heritability. In this study, the additive numerator relationship matrix **A** was estimated based on the marker genotype data for 402,509 SNPs, which consisted of the core SNPs (defined by the Rice SNP-Seek Database as the "404k CoreSNP Dataset") in all 12 chromosomes (this marker genotype data was prepared separately from the whole-genome sequence data), using the "A.mat" function in R package "rrBLUP" version 4.5 (Endelman and Jannink, 2012; Endelman, 2011).

The residual **e** in Eq. 6 was sampled identically and independently from the normal distribution, and was then normalized so that the narrow-sense heritability was equal to 0.6. Residual **e** was calculated using the following formula:

$$\mathbf{e} \sim \mathrm{MVN}(0, \mathbf{I}\sigma_{\mathrm{e}}^2), \qquad \text{(Eq. 6)}$$

where **I** is an identity matrix, and the residual variance σ_{e}^2 was determined so that the heritability was equal to 0.6.

Genome-wide association study (GWAS) using simulated data

We performed a GWAS on the seven non-mixture (A, B, C, D) and mixture populations (E, F, and G) using the marker genotype data and the simulated phenotypic data. We fitted the linear mixed model (Yu et al., 2006).

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{S}_i\alpha_i + \mathbf{Q\nu} + \mathbf{Zu} + \mathbf{e}, \qquad \text{(Eq. 7)}$$

where **y** is the vector of phenotypic values, $\mathbf{X\beta}$, $\mathbf{S}_i\alpha_i$, and $\mathbf{Q\nu}$ are the fixed effects terms, $\mathbf{Zu}$ is the random effects term, and **e** is the residuals vector. $\boldsymbol{\beta}$ represents all of the fixed effects other than $\mathbf{S}_i\boldsymbol{\alpha}_i$, and $\mathbf{Q}\boldsymbol{\nu}$, and X is the incidence design matrix corresponding to $\boldsymbol{\beta}$. In this study, $\mathbf{X\beta}$ was an intercept. $\mathbf{S}_i\alpha_i$ is composed of $\mathbf{S}_i$, which is the i_{th} marker of the genotype data, and α_i, which is the effect of that marker. $\mathbf{Q\nu}$ is the term used to correct for the effect of population structure, and in this study **Q** was the matrix of two eigenvectors corresponding to the upper two eigenvalues of the additive numerator for relationship matrix **A**, Finally, **u** represents the polygenetic effects, and **Z** is the incidence design matrix corresponding to **u**.

We used the EMMAX and P3D algorithms to reduce the computation time (Kennedy et al., 1992; Kang et al., 2008; 2010; Zhang et al., 2010). The "GWAS" function in R package "rrBLUP" version 4.5 (Endelman, 2011) was used to perform the GWAS described above.

Evaluation of the simulation results

The p-value (or $-\log_{10}(p)$) for each marker effect was estimated 100 times by the GWAS in five patterns according to the size of the F_{ST} for the seven non-mixture/mixture populations. In this study, the following summary statistics were mainly used to evaluate the GWAS results.

In the 100 simulations, the QTNs were not always polymorphic in each population (because the MAF of the whole population did not necessarily match the MAF of each individual population). In such cases, the $-\log_{10}(p)$ value of a QTN that was not polymorphic within a population could not be calculated. Therefore, when two SNPs were polymorphic within that population and were adjacent to the QTN, then the statistic of the more significant SNP was used as the QTN statistic. Since it was difficult to detect such QTNs using a GWAS, we calculated the summary statistics by dividing two patterns depending on polymorphism patterns of QTN1, i.e., whether using all simulation results or using only results whose QTN1 was polymorphic in the target population (referred to as "All" and "Polymorphic in the population", respectively).

253 **Correct detection rate (CDR) and $-\log_{10}(p)$**

254 The first summary statistic was whether the $-\log_{10}(p)$ rate for each QTN exceeded the
255 threshold in each GWAS (referred to as "CDR; correct detection rate"). We assumed that QTNs
256 would be successfully detected by the GWAS when the CDR was large. The $-\log_{10}(p)$ value
257 whose false discovery rate (FDR) was 0.05 was set as the threshold using the Benjamini-
258 Hochberg method (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). As the second
259 summary statistic, we used the $-\log_{10}(p)$ for each QTN in each GWAS, and we also assumed
260 that QTNs were successfully detected by the GWAS when this statistic was large.
261

262 **Area under the curve (AUC)**

263 We also regarded the mean of the AUC as one summary statistic. The AUC refers to the
264 area under the receiver operating characteristic (ROC) curve (Fig. S2 in Supplemental File 3),
265 which was obtained by plotting the false positive rate on the horizontal axis and the true positive
266 rate on the vertical axis when the threshold was varied (Hanley and McNeil, 1982). The AUC
267 was calculated using the following formula:

268 $$\mathrm{AUC} = \frac{1}{2} FPRs_1 TPRs_1 + \frac{1}{2} \sum_{i=2}^{m+1} (FPRs_i - FPRs_{i-1})(TPRs_i + TPRs_{i-1}), \qquad \text{(Eq. 8)}$$

269 where m is the number of QTNs, and $m = 3$ in this study. The $FPRs$ and $TPRs$ are the $m + 1$
270 vectors whose i_{th} elements are $FPRs_i$ and $TPRs_i$, respectively. $FPRs_i = TPRs_i = 1$ when $i =$
271 $m + 1$. When $1 \leqq i \leqq m$, the $FPRs_i$ and $TPRs_i$ represent the false positive rate and the true
272 positive rate at the time when i QTNs exceed the threshold, respectively. They were calculated
273 using the following formula:

$$FPRs_i = \frac{FP_i}{FP_i + TN_i} \qquad (1 \leqq i \leqq m), \qquad \text{(Eq. 9)}$$

$$TPRs_i = \frac{TP_i}{TP_i + FN_i} \qquad (1 \leqq i \leqq m), \qquad \text{(Eq. 10)}$$

where TP_i, FP_i, FN_i, and TN_i are the numbers of SNPs that are the true positives (where the SNP is a QTN and exceeds the threshold), the false positives (where the SNP is not a QTN but exceeds the threshold), the false negatives (where the SNP is a QTN but does not exceed the threshold), and the true negatives (where the SNP is not a QTN and does not exceed the threshold) at the time when i QTNs exceed the threshold respectively. When we evaluated the true/false positive rate, we considered the existence of linkage disequilibrium (LD) by investigating SNPs with LD as one set. In this study, we defined SNPs that satisfied the conditions that they were within 300 kb from the focused SNP and the condition that their squares of the correlation coefficients with the focused SNP were 0.35 or more as one set when considering LD. When we counted TP_i, FP_i, FN_i, and TN_i, we counted the number of the sets described above instead of the number of SNPs. The value for AUC calculated in this manner takes a value between 0 and 1. The GWAS is more successful when the AUC is closer to 1. Using the mean of the AUC as one of the summary statistics meant that it was possible to focus on each QTN and evaluate the overall results of the GWAS.

Precision, recall, and F-measure

We calculated the mean of precision, the mean of recall, and the mean of the *F*-measure as other summary statistics to evaluate the GWAS results. These summary statistics can be

295 calculated from the numbers of true positives, false positives, false negatives, and true negatives.
296 More specifically, the precision can be calculated using the following formula:

297 $$\text{Precision} = \frac{TP}{TP + FP}. \quad \text{(Eq. 11)}$$

298 We regarded an SNP as "positive" when the $-\log_{10}(p)$ of that SNP exceeded the
299 threshold described above. The precision represents the ratio of the detected SNPs that were
300 QTNs. The recall was defined using the following formula:

301 $$\text{Recall} = \frac{TP}{TP + FN}. \quad \text{(Eq. 12)}$$

302 The recall represents the proportion of QTNs detected by the GWAS. Finally, the *F*-
303 measure was calculated as the harmonic mean of the precision and the recall, and can be used to
304 comprehensively evaluate the GWAS results. The *F*-measure was calculated using the following
305 formula:

306 $$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad \text{(Eq. 13)}$$

307 The greater these summary statistics, the more accurate the GWAS results were.

308

309

310

311 **Degree of genetic diversity index**

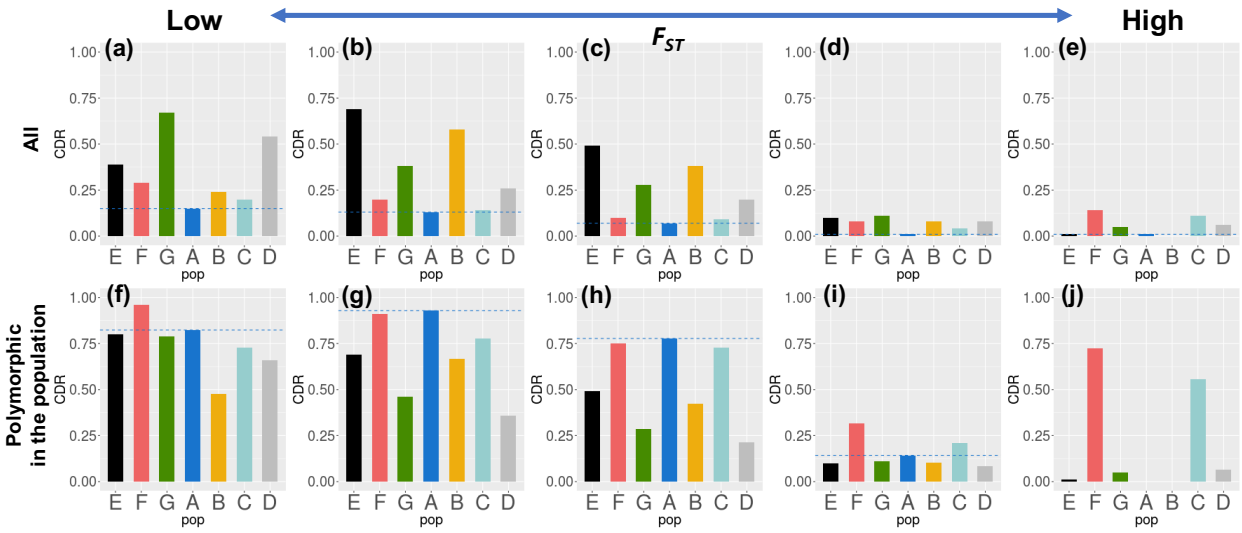
312 In order to evaluate the relationship between genetic diversity and the CDR results, we
313 prepared an index that indicated the degree of genetic diversity in each population. The

Euclidean distance matrix between accessions for each population was calculated. The median for the off-diagonal elements of the distance matrix was used to indicate the degree of genetic diversity (referred to as the "diversity level", Table 1). The median was chosen as the diversity level because the distribution of the distances between the accessions for E and G had a double peak. This was because, for mixture populations such as E and G, the distance within the subpopulations was small whereas the distance between subpopulations was large. Therefore, if the mean of the distances (almost the same as Nei's gene diversity index (NEI, 1973)) is chosen as the diversity level, then there is a risk of overestimating the diversity level.

RESULTS

Comparisons between the CDR and AUC for the QTN1s in each population

The CDRs of the QTN1s in each population were calculated under ten conditions: five levels of F_{ST} between A and B and two patterns of QTN polymorphism, i.e., whether the QTN was polymorphic or not in the target population (Fig. 2 and Table S2 in Supplemental File 5).



331

332 Fig. 2. **Correct detection rate for QTN1 in each population under ten conditions.**

333 The barplots of CDR of QTN1 in each population under ten conditions: five levels of F_{ST} of QTN1 and

334 two patterns of polymorphisms of QTN. Blue horizontal dashed lines indicate the CDR in the population

335 A for each population. A: *japonica*, B: *indica*, C: temperate *japonica*, D: diverse, E: A+ B, F: A + C, G:

336 A + D.

337

For almost all levels of F_{ST}, the CDRs for QTN1 in the mixture populations E, F, and G were larger than in the corresponding non-mixture populations B, C, and D, regardless of the two QTN polymorphism patterns (Fig. 2). The CDRs for QTN1 in the mixture populations E, F, and G were always larger than in population A when all the simulation results were taken into account (Fig. 2). When F_{ST} was low, and all simulation results were taken into account (Fig. 2a), populations G and D, which were highly diverse populations, had a higher CDR than the other populations. When F_{ST} was in the lower-middle or middle category, and all simulation results were taken into account (Figs. 2b, c), population E had the highest CDR. The CDR of the highly diverse populations G and D significantly decreased as F_{ST} increased. This result suggested that the QTN1 effect could confound with the population structure at higher F_{ST} values, which meant that it was difficult to detect QTN1 in a highly diverse population. When the F_{ST} value was in the higher-middle or high level categories, and all simulation results were taken into account, (Figs. 2d, e), the CDR for QTN1 became quite low in all populations. In populations D, E, and G, QTN1 was hardly detected because of the strong confounding effect of the population structure. In the other populations, the expected heterozygosity (H_e) for QTN1 was extremely small (In A and B, H_e was less than 0.1 in all 100 simulations). The small H_e may make the detection of QTN1 difficult.

We excluded the simulations in which there were no polymorphisms in the population so that the detection power of the GWAS when there were polymorphisms in an analyzed population could be evaluated (Figs. 2f-j). When F_{ST} was low, population F had the highest CDR and when F_{ST} was in the lower-middle or middle categories, population A had the highest CDR. However, there were only 14 and 9 cases in which QTN1 was polymorphic in population A. In general, the populations with low or moderate genetic diversities (A, C, and F) had higher CDRs

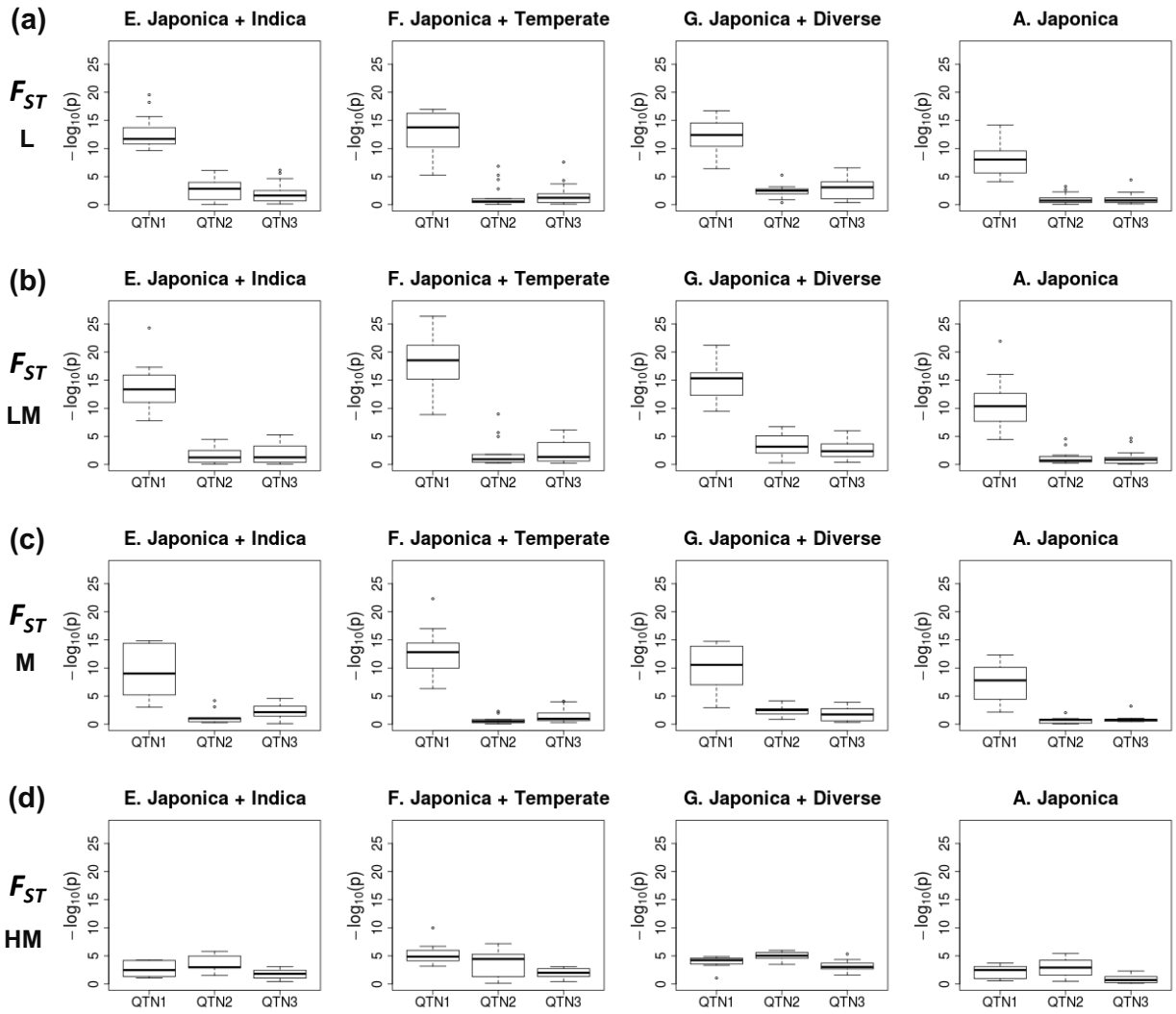
than the populations with high genetic diversities (D, E, and G). When F_{ST} was in the higher-middle or high categories, the results were similar to when F_{ST} was in the lower-middle or middle categories.

The CDRs of QTN2 and QTN3 were much lower than that of QTN1 because smaller genetic variances were assigned to these QTLs than QTN1 (Table S2). As in the case of QTN1, for almost all levels of F_{ST}, the CDRs of QTN2 and QTN3 were higher in the mixture populations (E, F, and G) than their corresponding non-mixture populations (B, C, and D). Furthermore, the CDRs for QTN2 and QTN3 in all the mixture populations were higher than for population A. The CDRs for QTN2 and QTN3 were also larger when the F_{ST} for QTN1 was higher.

Populations D and G had high AUC values in all cases (Table S2). Population F had a smaller AUC than populations D and G, even when the CDR was highest in population F.

Comparisons of the $-\log_{10}(p)$ values for the GWAS on each mixture population containing *japonica* (A)

We compared the $-\log_{10}(p)$ values for each QTN between populations mixed with the *japonica* population (A) to see if QTN1 was polymorphic in A (Fig. 3). Comparing these values allowed us to examine whether the detection power of the GWAS improved when genetic resources with higher genetic diversities were added to target population A. There is no plot for the high F_{ST} values because no QTN1 was polymorphic in population A over 100 simulations when the F_{ST} of QTN1 was high.



383

384 Fig. 3. **Boxplots of $-\log_{10}(p)$ of each QTN when QTN1 was polymorphic in *japonica* (A).**

385 Boxplots of $-\log_{10}(p)$ of each QTN for each mixture population and *japonica* (A) when QTN1

386 was polymorphic in A. These plots are shown divided into four categories according to the F_{ST}

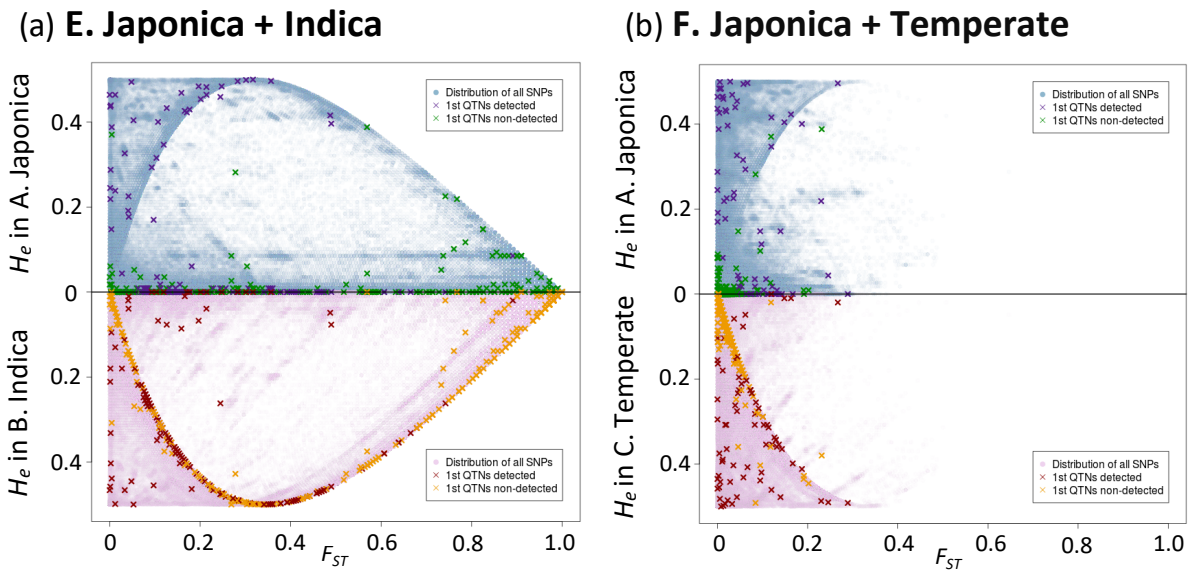
387 value for QTN1 (a: low, b: lower-middle, c: middle, d: higher-middle).

388

For all of the four F_{ST} levels, the detection power improved in all mixture populations compared to A (Fig. 3). Population F showed the highest detectability, and this tendency was conspicuous even when F_{ST} was in the middle or higher-middle categories (Figs. 3c, d, respectively). This is because the QTN1 effect is less likely to be confounded with the population structure in F than in the other mixture populations (E and G). Population G had the highest $-\log_{10}(p)$ values for QTN2 and QTN3, although only slightly (Fig. 3).

Factors affecting the detection power of QTNs in the mixture populations

We considered the factors related to the detection power of QTNs in the mixture populations by creating a figure that represented the relationship between F_{ST}, the expected heterozygosity (H_e), and the QTN1 detection power (Fig. 4 and Fig. S4 in Supplemental File 6).



402

403 Fig. 4. **Relationship between F_{ST}, H_e, and the detection power of QTN1.**

404 The distribution of each marker is plotted thinly with between subpopulation F_{ST} on the
405 horizontal axis and H_e of each subpopulation on the vertical axis. The dark X marks on the plot
406 show the SNPs selected as QTN1s in this study. Red and purple marks were detected by GWAS,
407 and green and yellow ones were not detected by GWAS.

408

409 Detection of the QTNs by the GWAS was generally difficult when the between-
410 subpopulation F_{ST} value was high, or H_e was low (Fig. 4). There seemed to be a significant
411 difference between plots F and E or G (Fig. 4a, Fig. S4, and Fig. 4b). However, in population F,
412 because A and C are genetically close, the F_{ST} between the subpopulations was not high.
413 Therefore, the relationships between F_{ST}, H_e, and the GWAS detection power applied to all
414 mixture populations.

415 Some of the QTNs were detected by the GWAS when F_{ST} was in the medium category, and
416 H_e in one of the subpopulations was close to 0 (Fig. 4a and Fig. S4). This suggested that even if
417 the QTN was fixed in one subpopulation, the QTN may still be detected by the GWAS if another
418 subpopulation was added to the analysis.

419

420

421

422

423 **Comparisons among the precision, recall, and *F*-measure values for each population**

424 The three summary statistics (the mean of precision, the mean of recall, and the mean of the
425 *F*-measure) were also calculated under ten conditions (Fig. S5 in Supplemental File 7). The
426 precision of the mixture populations was better than the precision value for population A for
427 almost all F_{ST} categories when all simulation results were taken into account (Fig. S5a). However,
428 it is not necessarily true that the precision of the mixture populations outperformed that of their
429 original genetic resources (compare E with B, F with C, and G with D). The recall values of the
430 mixture populations were larger than for their original genetic resources under all conditions.

Finally, a comparison of the *F*-measure for each population showed that there seemed to be no tendency associated with F_{ST}. Therefore, it was difficult to conclude which population was suitable for a GWAS when the *F*-measure is used. These results indicated that using mixture populations for a GWAS led to the detection of more SNPs, including QTNs.

Relationship between the CDR results and genetic diversity

The relationship between the CDR results for QTN1 and the degree of genetic diversity was evaluated under the two QTN polymorphism patterns, i.e., whether or not QTN was polymorphic in the population (Fig. S6 in Supplemental File 8). The CDRs for the mixture populations were usually larger than for the non-mixture populations if their diversity levels were close (Fig. S6a, b). A comparison of the results for the different F_{ST} categories showed that when F_{ST} was low, the populations with the highest diversities, such as D or G, had the highest CDRs, and when F_{ST} was in the lower-middle or middle categories, the populations with the second-highest diversities, such as B or E, had the highest CDRs. Finally, when F_{ST} was in the higher-middle or high categories, the populations with relatively low diversities, such as C or F, had the highest CDRs (Fig. S6a). However, when the simulations in which there were no polymorphisms in the population were excluded, the populations with relatively low diversities, such as A, C, or F, had the highest CDRs in almost all the F_{ST} categories (Fig. S6b).

DISCUSSION

Relationship between F_{ST} and QTN detection

One of the main results of this study was that the detection of QTNs was difficult in populations with high genetic diversities, such as D, E, and G, when the F_{ST} for QTN1 between *japonica* (A) and *indica* (B) was high. This was because the QTN effect confounds with the effect of population structure in these populations. We also examined the reasons why the CDRs for QTN2 and QTN3 were high when the QTN1 F_{ST} value was high.

In this study, phenotypic values were simulated using the following expression:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \mathbf{u} + \mathbf{e}, \quad \text{(Eq. 3)}$$

where $\mathbf{u}$ is the polygenetic effect, and is the term that reflects differences between accessions and thus differences between subpopulations. Therefore, if the degree of QTN1 genetic differentiation between *japonica* (A) and *indica* (B) is high, it can be assumed that there is a high correlation between $\mathbf{X}_1\beta_1$ and $\mathbf{u}$. In this study, we generated phenotypic values using a certain variance ratio under the assumption that each term is independent. Therefore, if there is a correlation between $\mathbf{X}_1\beta_1$ and $\mathbf{u}$, and the variance between these two terms is considered as one unit, it can be assumed that the variance is smaller than the total value of the two variances under the assumption of independence. Therefore, the variance of these two terms ($\mathbf{X}_1\beta_1 + \mathbf{u}$) in the total phenotypic variance becomes smaller, whereas the variances caused by the terms $\mathbf{X}_2\beta_2$ and $\mathbf{X}_3\beta_3$ become greater than those when it is assumed that each term is independent.

The GWAS model used in this study was

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{S}_i\alpha_i + \mathbf{Qv} + \mathbf{Zu} + \mathbf{e}, \quad \text{(Eq. 7)}$$

where $\mathbf{Q\nu}$ is the term used to correct the effect of population structure, and Zu shows the polygenetic effect. In this GWAS model, $\mathbf{S}_i\alpha_i$ and $\mathbf{Q\nu}$ or $\mathbf{Zu}$ have some correlation when $\mathbf{S}_i = \mathbf{X}_1$. This correlation results in the underestimation of α_i by the terms originally used to correct the effects of population structure or family relatedness, such as $\mathbf{Q\nu}$ and $\mathbf{Zu}$. Therefore, QTN detection is quite difficult when a GWAS is performed on mixture populations. For QTN2 and QTN3, where $\mathbf{S}_i = \mathbf{X}_2$ or $\mathbf{S}_i = \mathbf{X}_3$, there is generally no correlation between $\mathbf{S}_i\alpha_i$ and $\mathbf{Q\nu}$ or $\mathbf{Zu}$. Therefore, the detection of these QTNs is not related to these terms. Furthermore, the variances represented by the terms $\mathbf{X}_2\beta_2$ and $\mathbf{X}_3\beta_3$ are considered to be higher when the QTN1 genetic differentiation is not high. Therefore, the CDRs of QTN2 and QTN3 were high when the F_{ST} for QTN1 was high (Fig. 2 and Table S1). It has been suggested by Atwell et al. (2010) that a bias may occur in the GWAS results when the QTN correlates with population structure or family relatedness.

Relationship between H_e and the QTN detection

The detection of QTNs by a GWAS was difficult when the expected heterozygosity (H_e) in the population was low. When H_e in the population was low, the MAF was low, and alleles and mutations with low allele frequencies are known as "rare alleles" or "rare variants". In such cases, the QTN effect when the H_e values are low may be concealed by the QTN effect when H_e is not low or by the environmental effect because there are few accessions with one allele. For this reason, it is generally challenging to detect QTNs in such cases, but a method to deal with this problem has been developed (Wu et al., 2011).

One example of a “rare variant” that is common in plants is the haplotype condition. It has been reported that haplotypes are difficult to detect using a GWAS (Stram, 2014). This is because haplotypes are often “rare variants” and their H_e values in the population are low. Another problem caused by "rare variants" is that the non-causal SNP whose LD is strong with the “rare variant” may have a higher $-\log_{10}(p)$ value than the “rare variant”. This occurrence, known as “synthetic association”, often happens when the minor allele frequency of the SNP is higher than that of the “rare variant” (Dickson et al., 2010). These "synthetic associations" were often detected in this simulation study.

Summary and further discussion on each result

Generally, the CDRs of the QTNs showed that the populations suitable for a GWAS were different depending on whether all the QTNs were to be detected or only the polymorphic QTNs in the target population. Specifically, if all QTNs are to be detected when the degree of genetic differentiation between QTNs is low, then it is optimal to use a population with high genetic diversity that has as many polymorphisms as possible. However, as the degree of genetic differentiation becomes more extensive, a population with high genetic diversity is not suitable for a GWAS because the QTN effect is more likely to confound with the population structure. In contrast, a population with moderate genetic diversity, such as population F, was suitable for a GWAS, regardless of the degree of genetic differentiation. This was partly because the QTN1 effect was less likely to confound with the population structure in F than in E or G, even when F_{ST} was high. However, in either case, when the degree of genetic differentiation is extensive, it is difficult to detect the QTNs in any population. Therefore a GWAS analysis is not suitable,

518 which means that another approach, such as biparental QTL mapping, must be used to identify
519 genes (Lander and Botstein, 1989).

520 Population F had a smaller AUC than populations D and G, even when the CDR for
521 population F was the highest. From its definition, AUC is more dependent on how low
522 $-\log_{10}(p)$ of the QTN with the lowest $-\log_{10}(p)$ value is than on how high the $-\log_{10}(p)$ of
523 the QTN with the highest $-\log_{10}(p)$ value is. Furthermore, in this study, the number of markers
524 for the GWAS differed (Table 1). When $-\log_{10}(p)$ values for the QTNs were similar among the
525 different populations, the larger number of markers meant that the true negative rate increased,
526 and the false positive rate decreased in a population, which resulted in an increase in the AUC of
527 a population with a larger number of markers, e.g. D and G.

528 A comparison of the mixture populations and *japonica* (A) using $-\log_{10}(p)$ showed
529 that when the QTNs are polymorphic in a target population with low genetic diversity, genetic
530 resources with higher genetic diversities should be added to the target population. However, in
531 order to avoid cases where the degree of genetic differentiation among the QTNs is extensive
532 between the target population and genetic resources, it is desirable to use populations that are
533 genetically close to the target population.

534 Finally, the results suggested that the F_{ST} differences between the subpopulations and
535 the expected heterozygosity (H_e) of each subpopulation greatly influenced QTN detection by the
536 GWAS in the mixture populations (Fig. 4 and Fig. S4). This result was in agreement with the
537 above finding that QTN detection using a GWAS was generally difficult when F_{ST} was high, or
538 H_e were low. However, these situations frequently happened when the F_{ST} between the
539 subpopulations was moderate. Therefore, even if a QTN is fixed in one subpopulation, it may be

possible to detect the QTN by adding another population to the analysis because when the H_e of the QTN is low in one population and F_{ST} is moderate, it can be assumed that H_e is relatively high in the other population. Therefore, the H_e of the mixture population as a whole becomes larger and the detection of a QTN is possible unless the confounding of the effect of that QTN with the population structure is extensive. Although this situation is not difficult to interpret, it is extremely important that SNPs with high F_{ST} and low H_e values must exist in large numbers among populations. After taking this fact into account, a GWAS with a mixture population can be useful. Therefore, creating the proposed diagram shown in Fig. 4 and Fig. S4, will lead to a quantitative understanding of what kind of SNPs can be detected by a GWAS in mixture populations of interest.

Relationships with using whole-genome sequences

One of the major factors related to the QTN detection power was the fixation index F_{ST} differences among subpopulations. When the F_{ST} difference between the *japonica* (A) and *indica* (B) subpopulations was low, the CDR of the mixture populations was high. One example of such markers is that mutations may have occurred at the same position in both populations after they differentiated. Since such variants are relatively new variants, the LD relationship between these variants and surrounding markers will be weak. Therefore, these variants cannot be detected using marker genotype data with a small number of markers, such as an SNP array. However, the use of whole-genome sequences will increase the marker density, which improves the possibility of detecting such variants with a GWAS. In summary, using whole-genome sequences improves

562 the possibility of detecting QTNs with low F_{ST} values and the use of mixture populations should
563 further improve the QTN detection power. In this study, there were cases where SNPs in a low
564 LD region were selected as QTNs when F_{ST} was low.

565

566

567

568 **CONCLUSION**

569 In this study, we examined a way of selecting a population that was suitable for a GWAS by
570 conducting simulations using populations with various genetic backgrounds. We evaluated the
571 results of the simulations by dividing them into ten patterns according to two criteria: the degree
572 of genetic differentiation (F_{ST}) between two main subpopulations and QTN polymorphism in a
573 target population. When the QTNs are polymorphic in a target population, increasing the
574 population size by adding available genotypes to the target population improves the detection
575 power. We suggest that a population genetically similar to a target population is desirable. After
576 investigating F_{ST} and expected heterozygosity H_e as factors that may substantially influence the
577 detection power of a GWAS, the results showed that SNPs with high F_{ST} and low H_e values were
578 less likely to be detected by a GWAS that used mixture populations. These results indicated that
579 the detection power of a GWAS was improved by using mixture populations with different
580 genetic backgrounds. Furthermore, the use of publicly available whole-genome sequences meant
581 it was possible to increase the population size and to use polymorphic markers that were present
582 in high numbers, which should also improve the detection power of the GWAS.

583

584

585

586

587

588 ## ACKNOWLEDGMENTS

589 This study was supported by Grant-in-Aid for Scientific Research(A) (25252002), Grant-
590 in-Aid for Scientific Research(B) (Grant number 15H04436), JST, PRESTO (Grant number
591 JPMJPR15O6), the Cross-ministerial Strategic Innovation Promotion Program (SIP), the
592 "Technologies for creating next-generation agriculture, forestry and fisheries" (funding agency:
593 Bio-oriented Technology Research Advancement Institution, NARO), and JST CREST (Grant
594 Number JPMJCR16O).

595 We are also grateful for the advice given by Dr. Ryokei Tanaka.

596 ## SUPPLEMENTAL MATERIAL

597 Supplemental File 1: **Table S1.** Information about the 299 rice accessions used in this study.

598 Supplemental File 2: **Fig. S1.** Histogram showing the F_{ST} differences between *japonica* (A) and *indica*
599 (B).

600 Supplemental File 3: **Fig. S2.** Example of a ROC curve and the AUC.

601 Supplemental File 4: **Fig. S3.** Principal components analysis results for chromosome 1 and all the
602 chromosomes.

603 Supplemental File 5: **Table S2.** Correct detection rate rates for all QTNs and the AUC in each population.

Supplemental File 6: **Fig. S4.** Relationship between F_{ST}, H_e, and the QTN1 detection power for the population G.

Supplemental File 7: **Fig. S5.** Bar plots of the precision, the recall and the *F*-measure results.

Supplemental File 8: **Fig. S6.** Relationship between the diversity level and the CDR of QTN1.

Supplemental File 9: **Supplementary Note.** Additional information about the materials used in this study.

OPTIONAL SECTIONS

Availability of data and material

Whole genome sequencing data are available of 112 accessions of *Oryza sativa* subsp. *japonica* in the DDBJ Sequence Read Archive (SRA106223, ERA358140, DRA000158, DRA000307, DRA000897, DRA000927, DRA007273, DRA007256, and DRA008071). Whole genome sequencing data for all the other accessions are available in the "Rice SNP-Seek Database".

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KH, HKK, and HI conceived and designed the study. KH and HI performed the mathematical and statistical analysis. KH, HKK, MY, EK, SY and HN contributed to marker

genotyping. KH, HKK, and HI wrote the manuscript in consultation with MY, EK, SY, and HN.

All authors read and approved the final manuscript.

REFERENCES

Alexandrov, N., S. Tai, W. Wang, L. Mansueto, K. Palis, R.R. Fuentes, V.J. Ulat, et al. 2015. SNP-Seek Database of SNPs Derived from 3000 Rice Genomes. Nucleic Acids Res. 43(D1):D1023–27. doi: 10.1093/nar/gku1039.

Atwell, S., Y.S. Huang, B.J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, et al. 2010. Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines. Nature 465(7298):627–31. doi: 10.1038/nature08800.

Auwera, G.A. Van Der, M.O. Carneiro, C. Hartl, R. Poplin, A. Levy-moonshine, T. Jordan, K. Shakir, et al. 2014. From FastQ Data to High Confidence Varant Calls: The Genonme Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics 11(10):1–33. doi: 10.1002/0471250953.bi1110s43.From.

Begum, H., J.E. Spindel, A. Lalusin, T. Borromeo, G. Gregorio, J. Hernandez, P. Virk, B. Collard, and S.R. McCouch. 2015. Genome-Wide Association Mapping for Yield and Other Agronomic Traits in an Elite Breeding Population of Tropical Rice (Oryza Sativa). PLoS One 10(3):1–19. doi: 10.1371/journal.pone.0119873.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. 57(1):289–300. doi: 10.2307/2346101.

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data 30(15):2114–20. doi: 10.1093/bioinformatics/btu170.

Browning, B.L., and S.R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. 98(1):116–26. doi: 10.1016/j.ajhg.2015.11.020.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, et al. 2011. The Variant Call Format and VCFtools 27(15):2156–58. doi: 10.1093/bioinformatics/btr330.

Dickson, S.P., K. Wang, I. Krantz, H. Hakonarson, and D.B. Goldstein. 2010. Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biol. 8(1):e1000294. doi: 10.1371/journal.pbio.1000294.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package RrBLUP. Plant Genome J. 4(3):250. doi: 10.3835/plantgenome2011.08.0024.

Endelman, J.B., and J.L. Jannink. 2012. Shrinkage Estimation of the Realized Relationship Matrix. G3 (Bethesda) 2(11):1405–13. doi: 10.1534/g3.112.004259.

658 Hanley, J.A., and B.J. McNeil. 1982. The Meaning and Use of the Area under a Receiver
659 Operating Characteristic (ROC) Curve. Radiology 142(1):29–36. doi:
660 10.1080/02634938208400381.

661 Huang, X., Y. Zhao, X. Wei, C. Li, A. Wang, Q. Zhao, W. Li, et al. 2012. Genome-Wide
662 Association Study of Flowering Time and Grain Yield Traits in a Worldwide Collection of
663 Rice Germplasm. Nat. Genet. 44(1):32–39. doi: 10.1038/ng.1018.

664 Jarquin, D., H. Kajiya-Kanegae, C. Taishen, S. Yabe, R. Persa, J. Yu, H. Nakagawa, M.
665 Yamazaki, and H. Iwata. 2019. Coupling Day Length Data and Genomic Prediction Tools
666 for Predicting Time-Related Traits under Complex Scenarios, 703488.

667 Kang, H.M., J.H. Sul, S.K. Service, N.A. Zaitlen, S.Y. Kong, N.B. Freimer, C. Sabatti, and E.
668 Eskin. 2010. Variance Component Model to Account for Sample Structure in Genome-
669 Wide Association Studies. Nat. Genet. 42(4):348–54. doi: 10.1038/ng.548.

670 Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008.
671 Efficient Control of Population Structure in Model Organism Association Mapping.
672 Genetics 178(3):1709–23. doi: 10.1534/genetics.107.080101.

673 Kawahara, Y., M. de la Bastide, J.P. Hamilton, H. Kanamori, W.R. McCombie, S. Ouyang, D.C.
674 Schwartz, et al. 2013. Improvement of the Oryza Sativa Nipponbare Reference Genome
675 Using next Generation Sequence and Optical Map Data 6(1):1–10. doi: 10.1186/1939-8433-
676 6-1.

677 Kennedy, B.W., M. Quinton, and J.A. van Arendonk. 1992. Estimation of Effects of Single
678 Genes on Quantitative Traits. J. Anim. Sci. 70(7):2000–2012. doi:
679 10.1016/j.alcohol.2011.07.002.

680 Koboldt, D.C., K.M. Steinberg, D.E. Larson, R.K. Wilson, and E.R. Mardis. 2013. The Next-
681 Generation Sequencing Revolution and Its Impact on Genomics. Cell 155(1):27–38. doi:
682 10.1016/j.cell.2013.09.006.

683 Korte, A., and A. Farlow. 2013. The Advantages and Limitations of Trait Analysis with GWAS:
684 A Review. Plant Methods 9(1):29. doi: 10.1186/1746-4811-9-29.

685 Lander, E.S., and D. Botstein. 1989. Mapping Mendelian Factors Underlying Quantitative Traits
686 Using RFLP Linkage Maps. Genetics 121:185–99.
687 http://www.ncbi.nlm.nih.gov/pubmed/2563713.

688 Li, H. 2012. Exploring Single-Sample Snp and Indel Calling with Whole-Genome de Novo
689 Assembly 28(14):1838–44. doi: 10.1093/bioinformatics/bts280.

690 Li, J.Y., J. Wang, and R.S. Zeigler. 2014. The 3,000 Rice Genomes Project: New Opportunities
691 and Challenges for Future Rice Research. Gigascience 3(1):1–3. doi: 10.1186/2047-217X-
692 3-8.

693 Mansueto, L., R.R. Fuentes, F.N. Borja, J. Detras, J.M. Abrio-Santos, D. Chebotarov, M.
694 Sanciangco, et al. 2017. Rice SNP-Seek Database Update: New SNPs, Indels, and Queries.
695 Nucleic Acids Res. 45(D1):D1075–81. doi: 10.1093/nar/gkw1135.

696 Mansueto, L., R.R. Fuentes, D. Chebotarov, F.N. Borja, J. Detras, J.M. Abriol-Santos, K. Palis,
697 et al. 2016. SNP-Seek II: A Resource for Allele Mining and Analysis of Big Genomic Data
698 in Oryza Sativa. Curr. Plant Biol. 7–8:16–25. doi: 10.1016/j.cpb.2016.12.003.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2009. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. Proc. Int. Conf. Intellect. Capital, Knowl. Manag. Organ. Learn. 20:254–60. doi: 10.1101/gr.107524.110.20.

Metzker, M.L. 2010. Sequencing Technologies the next Generation. Nat. Rev. Genet. 11(1):31–46. doi: 10.1038/nrg2626.

Misra, G., S. Badoni, R. Anacleto, A. Graner, N. Alexandrov, and N. Sreenivasulu. 2017. Whole Genome Sequencing-Based Association Study to Unravel Genetic Architecture of Cooked Grain Width and Length Traits in Rice. Nat. Sci. Reports 7(1):12478. doi: 10.1038/s41598-017-12778-6.

NEI, M. 1973. Analysis of Gene Diversity in Subdivided Populations. Proc. Nat. Acad. Sci. USA 70(12):3321–23. doi: 10.1016/j.jasrep.2018.01.028.

Ott, J., J. Wang, and S.M. Leal. 2015. Genetic Linkage Analysis in the Age of Whole-Genome Sequencing. Nat. Rev. Genet. 16(5):275–84. doi: 10.1038/nrg3908.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R Language 20(2):289–90. doi: 10.1093/bioinformatics/btg412.

Saitou, N., and M. Nei. 1987. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4(4):406–25.

Storey, J.D., and R. Tibshirani. 2003. Statistical Significance for Genomewide Studies. Proc. Natl. Acad. Sci. 100(16):9440–45. doi: 10.1073/pnas.1530509100.

Stram, D. 2014. Design, Analysis, and Interpretation of Genome-Wide Association Scans. Heidelberg, New York: Springer Science+Business Media.

Wang, W., R. Mauleon, Z. Hu, D. Chebotarov, S. Tai, Z. Wu, M. Li, et al. 2018. Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice. Nature 557(7703):43–49. doi: 10.1038/s41586-018-0063-9.

Wright, S. 1965. THE INTERPRETATION OF POPULATION STRUCTURE BY F-STATISTICS WITH SPECIAL REGARD TO SYSTEMS OF MATING. Evolution (N. Y). 19:395–420.

Wu, M.C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am. J. Hum. Genet. 89(1):82–93. doi: 10.1016/j.ajhg.2011.05.029.

Yabe, S., M. Yamasaki, K. Ebana, T. Hayashi, and H. Iwata. 2016. Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops. PLoS One 11(4):1–21. doi: 10.1371/journal.pone.0153945.

Yano, K., E. Yamamoto, K. Aya, H. Takeuchi, P.C. Lo, L. Hu, M. Yamasaki, et al. 2016. Genome-Wide Association Study Using Whole-Genome Sequencing Rapidly Identifies New Genes Influencing Agronomic Traits in Rice. Nat. Genet. 48(8):927–34. doi: 10.1038/ng.3596.

737 Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, et al.
738 2006. A Unified Mixed-Model Method for Association Mapping That Accounts for
739 Multiple Levels of Relatedness. Nat. Genet. 38(2):203–8. doi: 10.1038/ng1702.

740 Zhang, Z., E. Ersoz, C.Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, et al.
741 2010. Mixed Linear Model Approach Adapted for Genome-Wide Association Studies. Nat.
742 Genet. 42(4):355–60. doi: 10.1038/ng.546.

743 Zhao, K., C.-W. Tung, G.C. Eizenga, M.H. Wright, M.L. Ali, A.H. Price, G.J. Norton, et al.
744 2011. Genome-Wide Association Mapping Reveals a Rich Genetic Architecture of
745 Complex Traits in Oryza Sativa. Nat. Commun. 2:467. doi: 10.1038/ncomms1467.

746

747 **FIGURES AND TABLES**

748 Fig. 1. **Unrooted phylogenetic tree plot for four non-mixture populations.**

749 Unrooted phylogenetic tree plot for the four non-mixture populations, which consisted of 112
750 accessions of *japonica* (A), 100 accessions of *indica* (B), 100 accessions of temperate *japonica*
751 (C), and 100 diverse accessions (D) with neighbor-joining method.

752

753 Fig. 2. **Correct detection rate for QTN1 in each population under ten conditions.**

754 The barplots of CDR of QTN1 in each population under ten conditions: five levels of F_{ST} of QTN1 and
755 two patterns of polymorphisms of QTN. Blue horizontal dashed lines indicate the CDR in the population
756 A for each population. A: *japonica*, B: *indica*, C: temperate *japonica*, D: diverse, E: A+ B, F: A + C, G:
757 A + D.

758

759 Fig. 3. **Boxplots of $-\log_{10}(p)$ of each QTN when QTN1 was polymorphic in *japonica* (A).**

760 Boxplots of $-\log_{10}(p)$ of each QTN for each mixture population and *japonica* (A) when QTN1
761 was polymorphic in A. These plots are shown divided into four categories according to the F_{ST}
762 value for QTN1 (a: low, b: lower-middle, c: middle, d: higher-middle).

763

764 Fig. 4. **Relationship between F_{ST}, H_e, and the detection power of QTN1.**

765 The distribution of each marker is plotted thinly with between subpopulation F_{ST} on the
766 horizontal axis and H_e of each subpopulation on the vertical axis. The dark X marks on the plot
767 show the SNPs selected as QTN1s in this study. Red and purple marks were detected by GWAS,
768 and green and yellow ones were not detected by GWAS.

769

770 **Table 1. Number of SNPs and the diversity level of non-mixture and mixture populations.**

	Population name	Number of accessions	Number of SNPs	Diversity level†
A.	Japonica	112	72,110	263.095
B.	Indica	100	427,943	660.416
C.	Temperate japonica	100	135,665	362.649
D.	Diverse	100	647,731	798.646
E.	A + B	212	633,507	803.064
F.	A + C	212	151,675	334.606
G.	A + D	212	684,774	859.678

771 † Diversity level is the index to indicate the degree of genetic diversity, which is described in the
772 Materials and Method section.