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

Kosuke Hamazaki1 | Hiromi Kajiya-Kanegae1,2 | Masanori Yamasaki3 | Kaworu Ebana4 | Shiori Yabe5 | Hiroshi Nakagawa6 | Hiroyoshi Iwata1

1Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The Univ. of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan
2Research Center for Agricultural Information Technology, National Agriculture and Food Research Organization, 3-1-1 Kannondai, Tsukuba, Ibaraki 305-8517, Japan
3Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe Univ., 1348 Uzurano, Kasai, Hyogo 675-2103, Japan
4Genetic Resources Center, National Agriculture and Food Research Organization, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan
5Institute of Crop Science, National Agriculture and Food Research Organization, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8518, Japan
6Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, 3-1-3 Kannondai, Tsukuba, Ibaraki 305-8604, Japan

Correspondence
Hiroyoshi Iwata, Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The Univ. of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.
Email: aiwata@mail.ecc.u-tokyo.ac.jp

Funding information
Scientific Research (A), Grant/Award Number: 25252002; Scientific Research (B), Grant/Award Number: 15H04436; JST Core Research for Evolutional Science and Technology, Grant/Award Number: JPMJCR16O; Technologies for creating next-generation agriculture, forestry and fisherie

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. Mixed populations containing subpopulations of different genetic backgrounds may be suitable populations. We conducted simulation experiments to see if a population with high genetic diversity, such as 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* L. subsp. *japonica*, mainly developed in Japan. We combined the target population with three populations that had higher genetic diversity. These were 100 *indica* accessions, 100 *japonica* accessions, and 100 accessions with various genetic backgrounds. The results showed that the GWAS’s power with a mixed population was generally higher than with a separate population. Also, the optimal GWAS populations varied depending on the fixation index ($F_{ST}$) of the quantitative trait nucleotides (QTNs) and the polymorphism of QTNs in each population. When a QTN was polymorphic in a target population, a target population combined with a higher diversity population improved the QTN’s detection power. By investigating $F_{ST}$ and the expected heterozygosity ($H_e$) as factors influencing the detection power, we

Abbreviations: AUC, area under the curve; CDR, correct detection rate; FN, false negative; FP, false positive; $F_{ST}$, fixation index; GWAS, genome-wide association study; $H_e$, expected heterozygosity; IND, *indica* population; LD, linkage disequilibrium; MAF, minor allele frequency; QTN, quantitative trait nucleotide; SNP, single nucleotide polymorphism; TJN, temperate *japonica* with a narrow genetic background; TN, true negative; TP, true positive.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *The Plant Genome* published by Wiley Periodicals, Inc. on behalf of Crop Science Society of America
showed that single nucleotide polymorphisms with high $F_{ST}$ or low $H_e$ are less likely to be detected by GWAS with mixed populations. Sequenced or genotyped germplasm collections can improve the GWAS’s detection power by using a subset of the collections with a target population.

1 | INTRODUCTION

Recently, as genome sequencing costs have continued to decrease (Metzker, 2010), the whole-genome sequences of a large number of cultivars or lines have become available for major crop species, such as rice (Li, Wang, & Zeigler, 2014; Wang et al., 2018). A GWAS based on whole-genome sequences can identify genes that control important agronomic traits more efficiently and accurately than previous methods (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013; Ott, Wang, & Leal, 2015; Yano et al., 2016).

It is important to prepare an appropriate population to be analyzed when attempting to detect candidate genes via GWAS techniques. For example, to avoid potential false positives caused by population stratification or structure, a GWAS population should be selected that results in low stratification (Begum et al., 2015; Yano et al., 2016). However, if such a population is selected as an analytical population for a GWAS, the sample size may be limited and the detection power of the GWAS will decrease (Korte & Farlow, 2013). Therefore, when designing an appropriate GWAS population, one should be aware of the tradeoff between population stratification and sample size.

When the population to be analyzed is being prepared, the factors that directly affect the GWAS 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 or lines have become publicly available through highly efficient sequencing analyses and database enrichment. The publicly available whole-genome sequence data will improve GWAS and could enable researchers to avoid the costs of sequencing analyses. For example, in rice, “The 3000 Rice Genomes Project” (Li et al., 2014; Wang et al., 2018) of 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 use existing public sequence data.

A mixed population obtained by mixing subpopulations with different genetic backgrounds could also potentially be used in a GWAS. One advantage of using such a mixed population is that it should improve the detection of causal variants by increasing the sample size. Conversely, a GWAS with a mixed 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 mixed population was carefully analyzed by a GWAS, with attention given to the confounding effects of the population structure.

An actual data analysis of rice using with-genome sequences showed that the detection power of a GWAS improved when $O.\text{ sativa subsp. japonica}$ and $O.\text{ sativa subsp. indica}$ populations were combined (Misra et al., 2017). Furthermore, the identification of new rice genes via 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 mixed populations or populations with high genetic diversity 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 mixed population in a GWAS. Furthermore, no previous studies have discussed which kinds of populations should be mixed to improve the GWAS’s 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 high genetic diversity compared with 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.

2 | MATERIALS AND METHODS

2.1 | Populations used in the GWAS

In this study, 112 accessions of $O.\text{ sativa subsp. japonica}$ [Population A, temperate $japonica$ with a narrow genetic background (TJN)] were used as a target population with low genetic diversity; these which were accessions that had mainly been developed in Japan (Yabe, Yamasaki, Ebana, Hayashi, & Iwata, 2016). We used the following three populations selected from “The 3000 Rice Genomes Project”
The Plant Genome (J. Y. Li et al., 2014), namely 100 accessions of *O. sativa* subsp. *indica* [Population B, *indica* population (IND)], 100 accessions of *Oryza sativa* subsp. *japonica* [Population C, temperate *japonica* with a wide genetic background (TJW)], and 100 accessions of *O. sativa* with various genetic backgrounds [Population D, diverse population (DIV)], as populations with higher diversities than the target population (Supplemental Table S1). The process used to select each of the 100 accessions is described in Supplemental File S1. One accession (IRIS ID: IRIS 313–11868) was duplicated in Populations B and D. Among Populations B, C, and D, the Population B was the most different from Population A, whereas Population C was the most similar to Population A. Population D contained *indica*, *japonica*, *aus*, and *aromatica* rice accessions, which meant that Population D had the highest genetic diversity. Figure 1 is an unrooted phylogenetic tree that shows the genetic relationships among accessions belonging to Populations A, B, C, and D.

The genetic relationships among the accessions were estimated by the neighbor-joining method (Saitou & Nei, 1987) in the R package “ape” version 5.3 (Paradis, Claude, & Strimmer, 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 mixed populations A + B, A + C, and A + D were named Populations E, F, and G, respectively. We compared the QTN detection power of the GWAS when the seven unmixed (A, B, C, and D) and mixed populations (E, F, and G) were used.

**Core Ideas**

- Genome-wide association studies with mixed populations are expected to improve detection power.
- The target population plus a diverse population improve quantitative trait nucleotides’ (QTNs’) detection power.
- The fixation index and the expected heterozygosities were related to detection power of QTNs.
- Germplasm collections are useful for improving the detection power of GWAS.

### 2.2 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 datasets deposited in the DNA Data Bank of Japan 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 by the Trimomatic version 0.36 program (Bolger, Lohse, & Usadel, 2014). The preprocessed reads were aligned with Os-Nipponbare-Reference-IRGSP-1.0 (Kawahara et al., 2013) and the bwa-0.7.12 mem algorithm with the default options (Li, 2012). The binary alignment map files deposited in the Rice SNP-Seek database were also reanalyzed. Single nucleotide polymorphism (SNP) calling was based on alignments determined by the Genome Analysis Toolkit version 3.7-0-gefedb67 (Van der Auwera et al., 2014; McKenna et al., 2009) and Picard package version 2.5.0 (http://broadinstitute.github.io/picard, accessed 17 Jan. 2020). The mapped reads were realigned with Realigner-TargetCreator and indelRealigner in the Genome Analysis Toolkit software. The SNPs and insertions and deletions were called at the population level by the UnifiedGenotyper in Genome Analysis Toolkit and the -glm BOTH option. We extracted biallelic sites in all the accessions from the variants with VCFtools version 0.1.13 (Danecek et al., 2011). Imputations were then imputed with Beagle version 4.1 (Browning & Browning, 2016). Finally, we analyzed the SNPs with minor allele frequencies (MAFs) of $\geq0.05$ in each population. In the analysis, the genotypes were represented as $-1$ (homozygous for the reference allele), 1 (homozygous for the alternative allele), or 0 (heterozygous for the reference and alternative alleles). Out of all the whole-genome sequence polymorphisms, only the SNPs on chromosome 1.
were analyzed. The number of SNPs on chromosome 1 in each population is shown in Table 1.

### 2.3 Generating phenotype data

Phenotypic data were simulated via the following formula:

\[
\mathbf{y} = \mathbf{X}_1 \mathbf{\beta}_1 + \mathbf{X}_2 \mathbf{\beta}_2 + \mathbf{X}_3 \mathbf{\beta}_3 + \mathbf{u} + \mathbf{e},
\]

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 QTNs with scores of -1, 0, or 1; \( \mathbf{\beta} = [\mathbf{\beta}_1 \mathbf{\beta}_2 \mathbf{\beta}_3]^T \) is the vector representing the effects of the three QTNs, \( \mathbf{u} \) is the vector for polygenetic effects, and \( \mathbf{e} \) is the residual 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, or high) on the basis of the \( F_{ST} \) between Populations A and B for the first QTN (Supplemental Figure S1). We assumed that the first QTN had four times more 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 pairwise \( F_{ST} \) between Populations 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:

\[
F_{ST} = 1 - \frac{H_S}{H_T},
\]

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:

\[
H_S = \frac{N_A \{2p_A (1 - p_A)\} + N_B \{2p_B (1 - p_B)\}}{N_A + N_B},
\]

\[
H_T = 2 \left( \frac{N_A p_A + N_B p_B}{N_A + N_B} \right) \left( 1 - \frac{N_A p_A + N_B p_B}{N_A + N_B} \right),
\]

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 Populations A and B is shown in Supplemental Figure S1, which also shows the thresholds for the five \( F_{ST} \) categories. We used the pairwise \( F_{ST} \) between Populations A and B as a summary statistic to represent the extent of the genetic differentiation of each SNP in the whole population.

The polygenetic effect in Equation 5 was sampled from the multivariate normal distribution whose variance-covariance matrix was proportional to the additive numerator relationship matrix \( \mathbf{A} \) and was normalized so that their variance was equal to that of the three QTN effects.

\[
\mathbf{u} \sim \text{MVN} (\mathbf{0}, \mathbf{G}).
\]

where \( \mathbf{G} = \mathbf{A} \sigma^2_A \) is the genetic covariance matrix; the additive genetic variance \( \sigma^2_A \) was automatically determined from the relationship with heritability. In this study, the additive numerator relationship matrix \( \mathbf{A} \) was estimated from 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 (these marker genotype data were prepared separately from the whole-genome sequence data), via the \( \mathbf{A} \).mat function in the R package rrBLUP version 4.5 (Endelman, 2011; Endelman & Jannink, 2012).

The residual \( \mathbf{e} \) in Equation 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. The residual \( \mathbf{e} \) was calculated via the following formula:

\[
\mathbf{e} \sim \text{MVN} (\mathbf{0}, \mathbf{I} \sigma^2_e),
\]

where \( \mathbf{I} \) is an identity matrix; the residual variance \( \sigma^2_e \) was determined so that the heritability was equal to 0.6.

In this study, we chose the experimental design and the parameter settings described above. This method was selected to clearly contrast the GWAS results between mixed and unmixed populations. For example, when the variance explained by the QTN of interest was lower than the choice because of a larger number of QTNs and/or lower heritability, it was difficult to evaluate the difference in the results caused by the genetic differentiation of the QTN. From this viewpoint, we focused on the five scenarios for \( F_{ST} \) as described...
above to evaluate how the genetic differentiation affected the GWAS results in a mixed or unmixed population.

### 2.4 | Genome-wide association study of simulated data

We performed a GWAS on the seven unmixed (Populations A, B, C, and D) and mixed populations (Populations E, F, and G) with the marker genotype data and the simulated phenotypic data. We fitted the linear mixed model (Yu et al., 2006), as shown in Equation 7:

\[
y = X\beta + S_i\alpha_i + Q\nu + Zu + e,
\]

where \(y\) is the vector of phenotypic values; \(X\beta\), \(S_i\alpha_i\), and \(Q\nu\) are the fixed effects terms; \(Zu\) is the random effect term; \(e\) is the residual vector; \(\beta\) represents all of the fixed effects other than \(S_i\alpha_i\) and \(Q\nu\); \(X\) is the incidence design matrix corresponding to \(\beta\). In this study, \(X\beta\) was an intercept. \(S_i\alpha_i\) is composed of \(S_i\), which is the \(i\)th marker of the genotype data, and \(\alpha_i\), which is the effect of that marker. \(Q\nu\) is the term used to correct for the effect of population structure; in this study, \(Q\) was the matrix of two eigenvectors corresponding to the upper two eigenvalues of the additive numerator for the relationship matrix \(A\). Finally, \(u\) represents the polygenic effects and \(Z\) is the incidence design matrix corresponding to \(u\).

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

### 2.5 | Evaluation of the simulation results

The \(p\)-value [or \(-\log_{10}(p)\) value] 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 unmixed and mixed 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, the statistic of the more significant SNP was used as the QTN statistic. Since it was difficult to detect such QTNs via GWAS, we calculated the summary statistics by dividing two patterns, depending on the polymorphism patterns of QTN1, namely whether we used all simulation results or using only results whose QTN1 was polymorphic in the target population.

### 2.6 | Correct detection rate (CDR) and \(-\log_{10}(p)\)

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

### 2.7 | Area under the curve

We also regarded the mean of the area under the curve (AUC) as one of the summary statistics. The AUC refers to the area under the receiver operating characteristic curve (Supplemental Figure S2), which was obtained by plotting the false positive rate on the horizontal axis and the true positive rate on the vertical axis when the threshold was varied (Hanley & McNeil, 1982). The AUC was calculated via the following formula:

\[
\text{AUC} = \frac{1}{2} FPR_i TPR_i + \frac{1}{2} \sum_{i=2}^{m+1} (FPR_i - FPR_{i-1}) (TPR_i + TPR_{i-1}),
\]

where \(m\) is the number of QTNs (\(m = 3\) in this study); \(FPR_i\) and \(TPR_i\) are the \(m + 1\) vectors whose \(i\)th elements are \(FPR_i\) and \(TPR_i\), respectively, where \(FPR_i = TPR_i = 1\) when \(i = m + 1\). When \(1 \leq i \leq m\), \(FPR_i\) and \(TPR_i\) represent the false positive rate and the true positive rate at the time when \(i\) QTNs exceed the threshold, respectively. They were calculated via the following formulae:

\[
FPR_i = \frac{FP_i}{FP_i + FN_i}, \quad (1 \leq i \leq m); \tag{9}
\]

\[
TPR_i = \frac{TP_i}{TP_i + FN_i}, \quad (1 \leq i \leq m), \tag{10}
\]

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 or 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 of being within 300 kb from the focused SNP and the condition that the squares of their correlation coefficients with the focused SNP were 0.35 or more as one set when LD was considered. When we counted $TP$, $FP$, $FN$, and $TN$, 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. Taking 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.

### 2.9 | Index of the degree of genetic diversity

In order to evaluate the relationship between genetic diversity and the CDR results, we 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 mixed 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.

### 3 | RESULTS

#### 3.1 | Comparisons between the CDR and AUC for the QTN1s in each population

The CDRs of the QTN1s in each population were calculated under 10 conditions: five levels of $F_{ST}$ between Populations A and B and two patterns of QTN polymorphism, namely whether the QTN was polymorphic or not in the target population (Figure 2, Supplemental Table S2).

For almost all levels of $F_{ST}$, the CDRs for QTN1 in the mixed populations (E, F, and G) were larger than in the corresponding unmixed populations (B, C, and D), regardless of the two QTN polymorphism patterns (Figure 2). The CDRs for QTN1 in the mixed populations (E, F, and G) were always larger than those in Population A when all the simulation results were taken into account (Figure 2). When $F_{ST}$ was low and all simulation results were taken into account (Figure 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 (Figure 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 categories and all simulation results were taken into account, (Figure 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 $H_c$ for QTN1
FIGURE 2 Correct detection rate for the first quantitative trait nucleotide (QTN1) in each population under 10 conditions

The barplots of the correct direction rate (CDR) of QTN1 in each population under 10 conditions: five levels of fixation index ($F_{ST}$) of QTN1 and two patterns of polymorphisms of QTN. Blue horizontal dashed lines indicate the CDR in the Population A for each population. (A) temperate japonica with a narrow genetic background (TJN, Population A); (B) indica population (IND, Population B); (C) temperate japonica with a wide genetic background (TJW, Population C); (D) Oryza sativa with diverse genetic backgrounds (DIV, Population D); (E) A + B; (F) A + C; (G) A + D was extremely small (In Populations 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 the polymorphisms were in a study population could be evaluated (Figure 2f–j). When $F_{ST}$ was low, population F had the highest CDR but 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 diversity (A, C, and F) had higher CDRs than the populations with high genetic diversity (D, E, and G). When $F_{ST}$ was in the higher-middle or high category, the results were similar to when $F_{ST}$ was in the lower-middle or middle category.

The CDRs of QTN2 and QTN3 were much lower than those of QTN1 because smaller genetic variances were assigned to these quantitative trait loci than QTN1 (Supplemental 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 mixed populations (E, F, and G) than in the corresponding unmixed populations (B, C, and D). Furthermore, the CDRs for QTN2 and QTN3 in all the mixed populations were higher than those 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 (Supplemental Table S2). Population F had a smaller AUC than Populations D and G, even when the CDR was highest in Population F.

3.2 | Comparisons of the $-\log_{10}(p)$ values

We compared the $-\log_{10}(p)$ values for each QTN between mixed populations and the TJN population (Population A) to see if QTN1 was polymorphic in Population A (Figure 3). Comparing these values allowed us to examine whether the detection power of the GWAS improved when genetic resources with higher genetic diversity were added to the 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.

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

3.3 | Factors affecting the detection power of QTNs in the mixed populations

We considered the factors related to the detection power of QTNs in the mixed populations by creating a figure that represented the relationship among $F_{ST}$, the $H_e$, and the QTN1 detection power (Figure 4 and Supplemental Figure S4).
FIGURE 3  Boxplots of the $-\log_{10}(p)$ of each quantitative trait nucleotide (QTN) when QTN1 was polymorphic in temperate japonica with a narrow genetic background

Boxplots of $-\log_{10}(p)$ of each QTN for each mixed population and temperate japonica with a narrow genetic background (TJN, Population A) when QTN1 was polymorphic in Population A. These plots are shown divided into four categories according to the fixation index ($F_{ST}$) value for QTN1 (a: low, b: lower-middle, c: middle, d: higher-middle).

Detection of the QTNs by the GWAS was generally difficult when the between-subpopulation $F_{ST}$ value was high or when $H_e$ was low (Figure 4). There seemed to be a significant difference between F (Figure 4b) and E (Figure 4a) or G (Supplemental Figure S4). However, in Population F, because Population A and C are genetically close, the $F_{ST}$ between the subpopulations was not high. Therefore, the relationships among $F_{ST}$, $H_e$, and the GWAS’s detection power applied to all mixed populations.

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

3.4 Comparisons among the precision, recall, and $F$-measure values for each population

The three summary statistics (the mean of precision, the mean of recall, and the mean of the $F$-measure) were also calculated under 10 conditions (Supplemental Figure S5). The precision of the mixed populations was better than the precision value for Population A for almost all $F_{ST}$ categories when all simulation results were taken into account (Supplemental...
3.5 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) (Supplemental Figure S6). The CDRs for the mixed populations were usually larger than for the unmixed populations if their diversity levels were close (Supplemental Figure 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 diversity, such as D or G, had the highest CDRs when $F_{ST}$ was in the lower-middle or middle category but the populations with the second-highest diversity, such as B or E, had the highest CDRs. Finally, when $F_{ST}$ was in the higher-middle or high category, the populations with relatively low diversity, such as C or F, had the highest CDRs (Supplemental Figure S6a). However, when the simulations in which there were no polymorphisms in the population were excluded, the populations with relatively low diversity, such as A, C, or F, had the highest CDRs in almost all the $F_{ST}$ categories (Supplemental Figure S6b).

4 DISCUSSION

4.1 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 diversity, such as D, E, and G, when the $F_{ST}$ for QTN1 between TJN (Population A) and IND (Population 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 via Equation (1), in which $\mathbf{u}$ is the polygenic effect and $\mathbf{X}_1\beta_1$ is the term that reflects differences between accessions and thus differences between subpopulations. Therefore, if the degree of QTN1 genetic differentiation between TJN (Population A) and IND (Population 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 with a certain variance ratio under the
assumption that each term is independent. Therefore, if there is a correlation between $X_i \beta_1$ and $u$ 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 ($X_i \beta_1 + u$) in the total phenotypic variance becomes smaller, whereas the variance caused by the terms $X_s \beta_2$ and $X_\beta_3$ becomes greater than those when it is assumed that each term is independent.

The GWAS model used in this study was as shown in Equation 7, in which $Qv$ is the term used to correct the effect of population structure and $Zu$ shows the polygenetic effect. In this GWAS model, $S_i \alpha_i$ and $Qv$ or $Zu$ have some correlation when $S_i = X_1$ ($S_i$ is the $i$th marker of the genotype data, $\alpha_i$ is the effect of that marker, and $X_1$ is the design matrix representing the genotypes QTN1). This correlation results in the underestimation of $\alpha_i$ by the terms originally used to correct the effects of population structure or family relatedness, such as $Qv$ and $Zu$. Therefore, QTN detection is quite difficult when a GWAS is performed on mixed populations. For QTN2 and QTN3, where $S_i = X_2$ or $S_i = X_3$, there is generally no correlation between $S_i \alpha_i$ and $Qv$ or $Zu$. Therefore, the detection of these QTNs is not related to these terms. Furthermore, the variances represented by the terms $X_s \beta_2$ and $X_\beta_3$ are considered to be higher when the genetic differentiation of QTN1 is not high. Therefore, the CDRs of QTN2 and QTN3 were high when the $F_{ST}$ for QTN1 was high (Figure 2 and Supplemental Table S1). It has been suggested by Atwell et al. (2010) that a bias may occur in the GWAS results when the QTN is correlated with population structure or family relatedness.

### 4.2 Relationship between $H_e$ and QTN detection

The detection of QTNs by a GWAS was difficult when the $H_e$ 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 to be rare alleles or rare variants. In such cases, the QTN effect when the $H_e$ values were low may be concealed by the QTN effect when $H_e$ was not low or by the environmental effect because there were few accessions with one allele. For this reason, it is generally challenging to detect QTNs in such cases, although a method to deal with this problem has been developed (Wu et al., 2011).

One problem caused by rare variants is that the noncausal SNP whose LD with the rare variant is strong may have a higher $-\log_{10}(p)$ value than the rare variant itself. This occurrence, known as synthetic association, often happens when the minor allele frequency of the noncausal SNP is higher than that of the rare variant (Dickson, Wang, Krantz, Hakonarson, & Goldstein, 2010). Synthetic associations were often detected in this simulation study.

### 4.3 Summary and further discussion of 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 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, which means that another approach, such as biparental quantitative trait locus mapping, must be used to identify genes (Lander & Botstein, 1989).

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

A comparison of the mixed populations and TJN (A) that applied $-\log_{10}(p)$ suggested that when the QTNs were polymorphic in a target population with low genetic diversity, genetic resources with higher genetic diversities should be added to the target population. However, in order to avoid cases where the degree of genetic differentiation among the QTNs is extensive between the target population and genetic resources, it is desirable to use populations that are genetically close to the target population.

The results also suggested that the $F_{ST}$ differences among the subpopulations and the $H_e$ of each subpopulation greatly influenced QTN detection by the GWAS in the mixed populations (Figure 4 and Supplemental Figure S4). This result was in agreement with the finding that detecting QTNs via GWAS was generally difficult when $F_{ST}$ was high or $H_e$ was low. However, these situations frequently happened when the $F_{ST}$
between the 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 mixed population as a whole becomes larger and the detection of a QTN is possible unless the confounding 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 mixed population can be useful. Therefore, creating the proposed diagram shown in Figure 4 and Supplemental Figure S6 will lead to a quantitative understanding of what kind of SNPs can be detected by a GWAS in mixed populations of interest.

Finally, we discuss a confounding effect between the diversity level and the absolute population size. In this study, we showed that QTNs could be detected more frequently in Populations E or G. In these cases, GWAS leveraged the diversity of these populations to detect more QTNs. Moreover, the power of GWAS in these populations was improved because of an increase in the population size. The influence of the two factors, however, can be distinguished by figures such as Supplemental Figure S6. For example, Populations C and F, and Populations D and E showed similar diversity levels but a mixed population with a larger population size, such as Populations F or E, showed a greater CDR.

4.4 | Comparison with the use of whole-genome sequences

One of the major factors related to the QTN detection power was the differences in $F_{ST}$ among subpopulations. When the $F_{ST}$ difference between the TJN (Population A) and IND (Population B) subpopulations was low, the CDR of the mixed 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 the surrounding markers will be weak. Therefore, these variants cannot be detected via 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, whole-genome sequences improve the possibility of detecting QTNs with low $F_{ST}$ values and the use of mixed populations should further improve the QTN detection power. In this study, there were cases where SNPs in a low LD region were selected as QTNs when $F_{ST}$ was low.

5 | CONCLUSION

In this study, we examined a way of selecting a population that was suitable for GWAS by conducting simulations of populations with various genetic backgrounds. We evaluated the results of the simulations by dividing them into 10 patterns according to two criteria: the degree of genetic differentiation ($F_{ST}$) between two main subpopulations and QTN polymorphism in a target population. When the QTNs were polymorphic in a target population, increasing the population size by adding available genotypes to the target population improved the detection power. We suggest that a population genetically similar to a target population is desirable. After we investigated $F_{ST}$ and $H_e$ as factors that may substantially influence the detection power of a GWAS, the results showed that SNPs with high $F_{ST}$ and low $H_e$ values were less likely to be detected by a GWAS that used mixed populations. These results indicated that the detection power of a GWAS was improved by using mixed populations with different genetic backgrounds. Furthermore, the use of publicly available whole-genome sequences meant it was possible to increase the population size and to use polymorphic markers that were present in high numbers, which should also have improved the detection power of the GWAS.

DATA AVAILABILITY

Whole-genome sequencing data of 112 accessions of *O. sativa* ssp. *japonica* have been deposited in the DNA Data Bank of Japan Sequence Read Archive (https://www.ddbj.nig.ac.jp/dra/index-e.html, accessed 24 Jan. 2020), under submission ID DRA008071 (http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA008071, accessed 24 Jan. 2020), SRA106223, ERA358140, DRA000158, DRA000307, DRA000897, DRA000927, DRA007273 and DRA007256. Whole-genome sequencing data for all the other accessions are available in the Rice SNP-Seek Database (https://snp-look.irri.org/_download.zul, accessed 24 Jan. 2020).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

KH, HKK, and HI conceived and designed the study. KH and HI performed the mathematical and statistical analyses. 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.
ACKNOWLEDGMENTS
This study was supported by Grant-in-Aid for Scientific Research (A) (25252002), Grant-in-Aid for Scientific Research (B) (grant number 15H04436), Japan Science and Technology Agency (JST), Precursory Research for Embryonic Science and Technology (grant number JPMPR1506), the Cross-Ministerial Strategic Innovation Promotion Program, the “Technologies for Creating Next-Generation Agriculture, Forestry and Fisheries” (funding agency: Bio-oriented Technology Research Advancement Institution, National Agriculture and Food Research Organization), and JST Core Research for Evolutional Science and Technology (grant number JPMICR160).

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

ORCID
Kosuke Hamazaki  https://orcid.org/0000-0002-7486-7438
Hiromi Kija-Kanegae  https://orcid.org/0000-0002-5719-7559
Masanori Yamasaki  https://orcid.org/0000-0001-7759-3965
Shiori Yabe  https://orcid.org/0000-0003-3200-3348
Hiroshi Nakagawa  https://orcid.org/0000-0001-9562-5991

REFERENCES
Alexandrov, N., Tai, S., Wang, W., Mansuetu, L., Palis, K., Fuentes, R. R., ... McNally, K. L. (2015). SNP-seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Research*, 43(D1), D1023–D1027. https://doi.org/10.1093/nar/gku1039.

Atwell, S., Huang, Y. S., Vilhjálmsdóttir, B. J., Willems, G., Horton, M., Li, Y., ... Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, 465(7329), 627–631. https://doi.org/10.1038/nature09800.

Begum, H., Spindel, J. E., Lalusis, A., Borrowe, T., Gregorio, G., Hernandez, J., ... McCouch, S. R. (2015). Genome-association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS One*, 10(3), 1–19. https://doi.org/10.1371/journal.pone.0119873.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *The Journal of the Royal Statistical Society*, 57(1), 289–300. https://doi.org/10.2307/2346101.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, 98(1), 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, 8(1), e1000294. https://doi.org/10.1371/journal.pbio.1000294.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package RfBLUP. *Plant Genome Journal*, 4(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

Endelman, J. B., & Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *G3 (Bethesda)*, 2(11), 1405–1413. https://doi.org/10.1534/g3.112.004259.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 142(1), 29–36. https://doi.org/10.1080/02634938208000381.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., ... Han, B. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, 44(1), 32–39. https://doi.org/10.1038/ng.1018.

Jarquin, D., Kajiy-Kanegae, H., Taishen, C., Yabe, S., Persa, R., Yu, J., ... Iwata, H. (2019). Coupling day length data and genomic prediction tools for predicting time-related traits under complex scenarios. Retrieved from https://www.biorxiv.org/content/10.1101/703488v3.full.

Jukes, T. H., & Cantor, C. R. (1969) Evolution of protein molecules. In H. N. Munro (Ed), *Mammalian protein metabolism* (pp. 21–132). New York, NY: Academic Press.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4), 348–354. https://doi.org/10.1038/ng.548.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3), 1709–1723. https://doi.org/10.1534/genetics.107.070101.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H.,McCombie, W. R., Ouyang, S., ... Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, 6, 4. https://doi.org/10.1186/1939-8433-6-4.

Kennedy, B. W., Quinton, M., & van Arendonk, J. A. (1992). Estimation of effects of single genes on quantitative traits. *Journal of Animal Science*, 70(7), 2000–2012. https://doi.org/10.1016/j.jalcohol.2011.07.002.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, 9(1), 29. https://doi.org/10.1186/1746-4811-9-29.

Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185–199.

Li, H. (2012). Exploring single-sample SNP and indel calling with whole-genome de novo assembly. *Bioinformatics*, 28(14), 1838–1844. https://doi.org/10.1093/bioinformatics/bts280.

Li, J. Y., Wang, J., & Zeigler, R. S. (2014). The 3,000 rice genomes project: New opportunities and challenges for future rice research. *Gigascience*, 3(1), 1–3. https://doi.org/10.1186/2047-217X-3-8.
Mansueto, L., Fuentes, R. R., Borja, F. N., Detras, J., Abriol-Santos, J. M., Chebotarow, D., … Alexandrov, N. (2017). Rice SNP-seek database update: New SNPs, indels, and queries. *Nucleic Acids Research*, 45(D1), D1075–81. https://doi.org/10.1093/nar/gkw1135.

Mansueto, L., Fuentes, R. R., Chebotarow, D., Borja, F. N., Detras, J., Abriol-Santos, J. M., … Mauleon, R. (2016). SNP-seek II: A resource for allele mining and analysis of big genomic data in *Oryza sativa*. *Current Plant Biology*, 7–8, 16–25. https://doi.org/10.1016/j.cpb.2016.12.003.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2009). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303. https://doi.org/10.1101/gr.107524.110.

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1), 31–46. https://doi.org/10.1038/nrg2626.

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., & Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Nature Science Reports*, 7(1), 12478. https://doi.org/10.1038/s41598-017-12778-6.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12), 3321–3323. https://doi.org/10.1016/j.crbp.2018.01.028.

Ott, J., Wang, J., & Leaf, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5), 275–284. https://doi.org/10.1038/nrg3908.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412.

Saitou, N., & Nei, M. (1987). The neighbor-joining method—A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Levy-Moonshine, A., Jordan, T., … DePristo, M. A. (2014). From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11(10), 1–33. https://doi.org/10.1002/0471250953.bi1110s43.

Wang, W., Mauleon, R., Hu, Z., Chebotarow, D., Tai, S., Wu, Z., … Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 557(7703), 43–49. https://doi.org/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., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029.

Yabe, S., Yamasaki, M., Ebana, K., Hayashi, T., & Iwata, H. (2016). Island-model genomic selection for long-term genetic improvement of autogamous crops. *PLoS One*, 11(4), 1–21. https://doi.org/10.1371/journal.pone.0153945.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., … Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, 48(8), 927–934. https://doi.org/10.1038/ng.3596.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., … Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2), 203–208. https://doi.org/10.1038/ng1702.

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., … Buckler, E.S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4), 355–360. https://doi.org/10.1038/ng.546.

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., … McCouch, S.R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications*, 2, 467. https://doi.org/10.1038/ncomms1467.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Hamazaki K, Kajiya-Kanegae H, Yamasaki M, et al. Choosing the optimal population for a genome-wide association study: A simulation of whole-genome sequences from rice. *Plant Genome*. 2020;13:e20005. https://doi.org/10.1002/tpg2.20005