Designing a Mini-Core Collection Effectively Representing 3004 Diverse Rice Accessions

Angad Kumar1,2, Shivendra Kumar1,2, Kajol B.M. Singh1, Manoj Prasad1 and Jitendra K. Thakur1,*

1Plant Mediator Lab, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India
2These authors contributed equally to this article.
*Correspondence: Jitendra K. Thakur (jthakur@nipgr.ac.in)
https://doi.org/10.1016/j.xplc.2020.100049

ABSTRACT

Genetic diversity provides the foundation for plant breeding and genetic research. Over 3000 rice genomes were recently sequenced as part of the 3K Rice Genome (3KRG) Project. We added four additional Indian rice accessions to create a panel of 3004 accessions. However, such a large collection of germplasm is difficult to preserve and evaluate. The construction of core and mini-core collections is an efficient method for the management of genetic resources. In this study, we developed a mini-core comprising 520 accessions that captured most of the SNPs and represented all of the phenotypes and geographic regions from the original panel. The mini-core was validated using different statistical analyses and contained representatives from all major rice groups, including japonica, indica, aus/boro, and aromatic/basmati. Genome-wide association analyses of the mini-core panel efficiently reproduced the marker–trait associations identified in the original panel. Haplotype analysis validated the utility of the mini-core panel. In the current era with many ongoing large-scale sequencing projects, such a strategy for mini-core design should be useful in many crops. The rice mini-core collection developed in this study would be valuable for agronomic trait evaluation and useful for rice improvement via marker-assisted molecular breeding.

Keywords: rice, mini-core, SNPs, GWAS, 3KRG, agronomic trait

Kumar A., Kumar S., Singh K.B.M., Prasad M., and Thakur J.K. (2020). Designing a Mini-Core Collection Effectively Representing 3004 Diverse Rice Accessions. Plant Comm. 1, 100049.

INTRODUCTION

Rice (Oryza sativa) is among the primary staple crops that fulfill the nutritional requirements of more than half of the world’s population. Improvements in global rice production will have a direct impact on meeting the world’s growing food demand. More than 90% of global rice production is contributed by Asian countries, particularly China and India (FAOSTAT, 2017). India is the second largest producer of rice (165.3 million tons) after China (208.4 million tons) and accounts for ~22% of total global rice production. Increases in rice yield are achieved mainly through improved cropping methods, fertilizer use, and—in many areas—intensive irrigation. However, the outcome of these strategies is now reaching saturation and becoming limited, and there is a demand for alternative means of yield improvement. The genetic improvement of rice cultivars and varieties can be an effective strategy in this regard. Yield improvement can be realized through breeding programs that incorporate marker-assisted selection and genetic methods to identify new sources of genetic variation that may help to increase productivity (McCouch et al., 2016). Productivity/yield is a complex trait that is governed by multiple genes and depends on both genetic composition and environmental factors. Variability arises due to segregating alleles at multiple loci whose individual effects on the phenotypic trait are relatively small, and the overall expression is also influenced by environmental conditions. Single-nucleotide polymorphisms (SNPs), present throughout the genome, are one of the major causes of allelic variation that underlie genetic variability in a population. Genetic variation leads to a multitude of phenotypes, which form the basis for selection of improved cultivars for breeding and agricultural purposes. Identification of loci that govern quantitative traits is critical for the maintenance of variation within and among populations. Identification of quantitative trait loci (QTLs) by the conventional method of linkage mapping or QTL mapping involves the development of a mapping population, a time-consuming process that captures a limited number of recombination events based on parental combinations. This methodology forms a part of the marker-
Plant Communications

assisted selection and biotechnological approach that has been used in a large number of crops for the identification of genes that govern complex traits (Edgerton, 2009; Morrell et al., 2012).

With advances in high-throughput genome sequencing and phenotyping methods, genome-wide association studies (GWAS) have been initiated. GWAS analysis has proved to be very effective for crop improvement. It is a very efficient approach for the identification of marker–trait associations and has been used to identify genes or loci that govern complex traits (Gupta et al., 2005; Breseghello and Sorrells, 2006; Huang et al., 2010, 2012; Famoso et al., 2011; Ingvarsson and Street, 2011; Kump et al., 2011; Zhao et al., 2011; Morrell et al., 2012). The advantage of GWAS is that it does not require a mapping population. It explores the genomic and phenotypic diversity present in the available collection to assess marker–trait associations. It also captures a large number of historical recombination events that are prevalent in the population. The basic requirement for GWAS is a diverse panel that harbors historical recombination events for greater genetic resolution (Morrell et al., 2012). This purpose is best served by a core collection that is designed to capture the maximum available/diversity (genetic, phenotypic, and geographic) of the entire population, with a limited number of individuals that share low or no kinship (Korte et al., 2012). Core collections have been used as association panels for GWAS in different studies (El Bakkali et al., 2013; Zhang et al., 2014; Perseguini et al., 2015; Ambreen et al., 2018). In the case of rice, attempts have been made to generate core collections and use them as association panels. The US Department of Agriculture MC collection consists of 217 accessions that represent the genotypic and phenotypic diversity of the rice core subset of 1794 accessions, but it is based on a small number of simple sequence repeats (SSRs) and InDel markers (Agrama et al., 2009). More recently, a Rice Diversity Panel was developed that consisted of different collections: Rice Diversity Panel 1 (RDP1), Rice Diversity Panel 2 (RDP2), and a collection from the Institute of Agrobiological Sciences, NARO (Eizenga et al., 2014; Ebana et al., 2008; McCouch et al., 2016). However, the accessions in these panels were genotyped with a fixed array of 700K SNPs (Liakat Ali et al., 2011; Eizenga et al., 2014; McCouch et al., 2016).

Recently, with the availability of a resequencing dataset for 3000 diverse rice accessions that generated 32 million SNPs, a deep and robust platform has been provided to promote marker–trait associations and has been used to identify genes or loci that govern complex traits (Gupta et al., 2005; Breseghello and Sorrells, 2006; Huang et al., 2010, 2012; Famoso et al., 2011; Ingvarsson and Street, 2011; Kump et al., 2011; Zhao et al., 2011; Morrell et al., 2012). The advantage of GWAS is that it does not require a mapping population. It explores the genomic and phenotypic diversity present in the available collection to assess marker–trait associations. It also captures a large number of historical recombination events that are prevalent in the population. The basic requirement for GWAS is a diverse panel that harbors historical recombination events for greater genetic resolution (Morrell et al., 2012). This purpose is best served by a core collection that is designed to capture the maximum available/diversity (genetic, phenotypic, and geographic) of the entire population, with a limited number of individuals that share low or no kinship (Korte et al., 2012). Core collections have been used as association panels for GWAS in different studies (El Bakkali et al., 2013; Zhang et al., 2014; Perseguini et al., 2015; Ambreen et al., 2018). In the case of rice, attempts have been made to generate core collections and use them as association panels. The US Department of Agriculture MC collection consists of 217 accessions that represent the genotypic and phenotypic diversity of the rice core subset of 1794 accessions, but it is based on a small number of simple sequence repeats (SSRs) and InDel markers (Agrama et al., 2009). More recently, a Rice Diversity Panel was developed that consisted of different collections: Rice Diversity Panel 1 (RDP1), Rice Diversity Panel 2 (RDP2), and a collection from the Institute of Agrobiological Sciences, NARO (Eizenga et al., 2014; Ebana et al., 2008; McCouch et al., 2016). However, the accessions in these panels were genotyped with a fixed array of 700K SNPs (Liakat Ali et al., 2011; Eizenga et al., 2014; McCouch et al., 2016).

RESULTS AND DISCUSSION

Generation of the SNP Dataset

We used publically available SNP data from the 3000 rice accessions in the 3K Rice Genome (3KRG) project (Alexandrov et al., 2015; Mansueto et al., 2017). In addition, we resequenced four Indian rice accessions: LGR, PB 1121, Sonasal, and Bindli. After filtering and alignment, we obtained 3 564 117 high-quality SNPs for these four genotypes with reference to the Nipponbare genome. The SNP read depth varied from 10 to more than 11 000, and the overall sequencing depth for the four rice accessions ranged from 42× to 48×. In the present study, we combined the new Indian rice dataset with the 3KRG SNP dataset. Overall, 18.9 million SNPs were identified among the 3000 sequenced genomes with an average depth of ~14×, ranging from ~4× to 60×. To bring the 3KRG dataset to the same level of quality as the new data, we considered the filtered dataset (~4 800 000 SNPs) corrected for excess of heterozygosity and linkage disequilibrium (LD). Finally, we merged both datasets and identified their common SNPs (2 081 521). The common SNPs were non-uniformly distributed over different rice chromosomes. The greatest number of SNPs were located on chromosomes 1, 11, and 2, whereas the smallest number of SNPs was found on chromosome 9.

Development of the Mini-Core Collection

To create a representative mini-core group, we used genotypic data (SNPs) from 3000 rice accessions and phenotypic data on 18 agronomic traits from 2266 rice accessions (Mansueto et al., 2017). We initially chose to develop independent mini-cores from phenotypic data and genotypic data to avoid tradeoffs and capture the maximum possible phenotypic and genotypic variability present in the original collection. For the phenotype-based subset, scanning of 2266 accessions resulted in a mini-core collection (CC1) of 227 accessions that represented 10% of the initial collection. We added the four Indian accessions (LGR, PB 1121, Sonasal, and Bindli) to this panel because of their notable genomic and phenotypic diversity. Mini-core collection CC1 therefore consisted of 231 accessions representing diversity in phenotypic traits. The 3000 accessions with their SNP data were analyzed separately for the development of a second mini-core collection (CC2) consisting of 300 accessions that represented 10% of the original collection and also included the four Indian accessions sequenced in our laboratory.

Mini-core collections CC1 and CC2 were assessed for their coverage of phenotypic variation with reference to the original panel (Supplemental Table 1). Neither of the two mini-cores captured the entire range of phenotypic traits present in the original collection. Traits that could not be captured in the mini-cores were evaluated with reference to the original panel.
Design of a Rice Mini-Core for Association Studies

Figure 1. Venn Diagram Showing the Distribution of Accessions in Different Mini-Core Collections Developed in This Study.

CC1 represents the mini-core designed using phenotypic data. CC2 represents the mini-core designed using SNP data. CC3 represents the merged (CC1 + CC2 + 17 accessions) mini-core collection.

The mini-core collections were also assessed for their representation of all the varietal groups and regional gene pools present in the original panel (Supplemental Tables 3 and 4). The most prevalent group in mini-core CC1 was indica (129 accessions), followed by Temperate japonica (38), Intermediate (19), Tropical japonica (15), japonica (14), aus/boro (11), and Aromatic (5). The most prevalent group in mini-core CC2 was indica (171), followed by Intermediate (45), aus/boro (42), Aromatic and japonica (12 each), Tropical japonica (10), and Temperate japonica (8) (Supplemental Table 3). We also compared the distribution of accessions from different varietal groups in the mini-cores and found that CC2 had a higher proportion of accessions from the aus/boro (19.5% of the original collection), Intermediate (33.3%), and Aromatic (16.9%) groups. On the other hand, CC1 had only 5.1% of the original representation from aus/boro, 14% from Intermediate, and 7% from Aromatic (Supplemental Table 3). Accessions from the Temperate japonica group were highly represented in CC1 (11.9% of the original representation), whereas only a small proportion (2.5%) was represented in CC2 (Supplemental Table 3). Comparable portions of accessions from the indica (7.4% in CC1 and 9.8% in CC2), Tropical japonica (3.8% in CC1 and 2.5% in CC2), and japonica (10.6% in CC1 and 9% in CC2) groups were present in both CC1 and CC2 mini-cores. Thus, neither of the two mini-cores developed here contained 10% of the representatives from all varietal groups (Supplemental Table 3).

The mini-core collections were then assessed for their distribution of accessions from different regional gene pools (Supplemental Table 4). Mini-core CC1 contained 55 accessions from South Asia (6.9% of the original collection), followed by 52 accessions from South East Asia (5.1%), 52 accessions from China (10.8%), 18 accessions from Europe (15.2%), 17 accessions from America (10.2%), 15 accessions each from East Asia and Africa (11.4% and 5.9%, respectively), four accessions from Oceania (23.5%), and three accessions of unknown origin (8.8%). CC2 contained 122 accessions from South Asia (15.5% of the original collection), followed by 70 accessions from South East Asia (6.9%), 55 accessions from China (11.4%), 23 accessions from Africa (6.1%), 13 accessions from America (7.8%), eight accessions from East Asia (6%), six accessions of unknown origin (17.4%), two accessions from Europe (1%), and one accession from Oceania (5.9%; Supplemental Table 4).

Only 28 accessions were shared between CC1 and CC2, showing that different accessions were selected on the basis of phenotypic and genotypic variation and justifying our concern about designing the mini-cores independently using only phenotypic or genotypic data. An ideal mini-core should represent the maximum possible diversity present in the original collection. However, different evaluation criteria such as phenotypic range (Supplemental Table 1), MD%, VD%, VR%, CR%, Shannon’s and Nei’s indices (Supplemental Table 2), and varietal (Supplemental Table 3) and geographic coverage (Supplemental Table 4) revealed that neither of the mini-cores (CC1 and CC2) captured sufficient diversity from the original collection to be considered an ideal representative subset. Therefore, we merged CC1 and CC2 to develop mini-core collection CC3, comprising 520 non-redundant accessions (503 accessions from the merging of CC1 and CC2 and 17 accessions that captured the extreme values of the phenotypic traits discussed below), in order to capture the maximum possible allele/trait diversity and prevent any tradeoffs between the two datasets (phenotypic and genotypic) when used in conjunction (Figure 1). The 520 accessions of CC3 represented 17.3% of the original collection (3004 accessions) and fulfilled the initial size requirement for an ideal core collection, which should range between 5% and 20% of the original collection (Brown and Spilllane, 1999). CC3 was assessed for its representation of the original collection and various traits under consideration by different evaluation criteria (Supplemental Tables 1–4). CC3 covered the entire range of traits from the original collection, including traits not completely covered by CC1 and CC2, such as days to 80% heading (DEH), 100 grain weight (HGW), days to first flowering (DFF), grain width (GW), panicle length (PL), and seedling height (SH). The two mini-core collections were further assessed with various evaluation criteria such as Shannon’s diversity index, Nei’s gene diversity, mean difference percentage (MD%), variance difference percentage (VD%), variable rate of coefficient of variance (VR%), and coincidence rate of range (CR%) to assess their efficiency in capturing the maximum diversity present in the original collection. The MD% of the mini-core collections ranged from 2.8% to 4.08%, well below the prescribed value of 20% (Supplemental Table 2). VD%, which represents the variance captured in the mini-core collections, ranged from 19.78% to 39.77%. The VR% ranged from 86% for CC1 to 107.68% for CC2. CC1 had a higher CR% value of 92%, whereas that of CC2 was 91.1%. The value of the Shannon–Weaver index (H) ranged from 1.98 for CC1 to 2.25 for CC2. The value of Nei’s genetic diversity (I) was higher for CC2 (0.79) than for CC1 (0.77) (Supplemental Table 2).
**Plant Communications**

![Figure 2. Grouping of the 3004 Rice Accessions Based on Polymorphic SNP Markers.](image)

(A) Maximum-likelihood dendrogram illustrating the genetic relationships among accessions. The two clusters were designated CL I (CL Ia, b) and CL II (CL IIa–c) with further subclustering shown.

(B) Principal component analysis of the 3004 accessions from the original collection, showing principal component axes 1 and 2. The distribution of accessions in different quadrants (I–IV) is shown. Varietal group color codes are provided. Color codes representing different varietal groups are given on the right.

Design of a Rice Mini-Core for Association Studies

The number of accessions and the percentage from the original collection (27 accessions; Supplemental Tables 3 and 4) were evenly distributed along coordinate axes 1 and 2, which accounted for 45.6% and 26% of the total variance, respectively (Figure 2B). The *indica* accessions clustered together in the PCA, consistent with the results of the distance-based analysis. They formed the largest group in the original collection and were mainly present in Cluster Ia of the distance-based analysis and quadrants I and IV of the PCA. The *japonica*, Temperate *japonica*, and Tropical *japonica* accessions were part of Cluster II in the distance-based analysis and were found in quadrants III

**Table 3.** The number of accessions and the percentage representation of all varietal groups from the original collection in mini-core CC3 are presented in Supplemental Table 3. Representation of accessions from different regional gene pools in the original collection varied from 12% to 29.4% in mini-core CC3 (Supplemental Table 4). The most prevalent region in CC3 was South Asia (176 accessions; 22.4% of its original representation), followed by China (101 accessions; 20.95%), Africa (38 accessions; 15%), America (28 accessions; 16.9%), East Asia (21 accessions; 15.9%), Europe (19 accessions; 16.1%), unknown origin (9 accessions; 25.5%), and Oceania (5 accessions; 29.4%; Supplemental Table 4). Thus, mini-core collection CC3 more successfully fulfilled the criteria for capturing the maximum possible diversity from the original panel than did mini-cores CC1 and CC2, and it was considered further for its utility as an association panel.

**Distance-Based Cluster Analysis and Principal Component Analysis**

Distance-based cluster analysis was performed to assess the grouping of accessions from the original collection of rice genotypes. Analysis of the SNP data (2 081 521 SNPs) using the maximum-likelihood method grouped the 3004 accessions into two major clusters (CL I and CL II) with internal subgroupings (Figure 2A). The CL I cluster contained the greatest number of accessions (66%) from the original collection, and CL II contained approximately 32% of the original accessions (Supplemental Table 5). Approximately 1.4% of the accessions did not belong to either of the clusters (Figure 2A). These unclustered accessions (43) were mainly Intermediate (21) and *indica* (15) genotypes. In addition, some of the *japonica* (3), Tropical *japonica* (2), Temperate *japonica* (1), and Aromatic (1) genotypes also remained unclustered. The 1987 accessions of Cluster CL I were further grouped into two subclusters, CL Ia and CL Ib. The larger subcluster, CL Ia, consisted of 1771 accessions and was mainly dominated by *indica* (1641) genotypes, whereas cluster CL Ib consisted of 216 accessions with major contributions from *aus/boro* (172) and *indica* (25) genotypes. The 974 accessions of CL II were divided into three subclusters, CL IIa, IIb, and IIc. The largest subcluster, CL IIa, contained 519 accessions and was mainly dominated by Tropical *japonica* (329) and *japonica* (80) genotypes. Subcluster CL IIb contained 358 accessions and was dominated by Temperate *japonica* (250) genotypes. The Indian genotype LGR was also part of subcluster CL IIb. CL IIc was the smallest subcluster of CL II and contained 97 accessions, with major representation from Aromatic (50) and Intermediate (23) genotypes. Notably, Bindli, PB 1121, and Sonasal grouped together in subcluster CL IIc (Figure 2A and Supplemental Table 5).

In the principal component analysis (PCA), the 3004 original accessions were evenly distributed along coordinate axes 1 and 2, which accounted for 45.6% and 26% of the total variance, respectively (Figure 2B). The *indica* accessions clustered together in the PCA, consistent with the results of the distance-based analysis. They formed the largest group in the original collection and were mainly present in Cluster Ia of the distance-based analysis and quadrants I and IV of the PCA. The *japonica*, Temperate *japonica*, and Tropical *japonica* accessions were part of Cluster II in the distance-based analysis and were found in quadrants III...
of the PCA were present in CC3 (Figure 3B). Thus, CC3 had contained representative accessions from all clusters of the maximum-likelihood dendrogram and all quadrants of the PCA, capturing the maximum possible genotypic diversity.

Population Structure Analysis of the Original Collection Using FastSTRUCTURE

Population structure analysis of the original 3004 rice accessions was performed using the FastSTRUCTURE program (Raj et al., 2014). The best clustering was observed at $K = 7$, and the clusters obtained were named FSTR CL 1–7 (Figure 4A). FSTR CL 1 consisted of 219 accessions and was mainly dominated by *aus/boro* (179 accessions) and *indica* genotypes (28; Supplemental Table 7); it showed congruence with CL Ib from the maximum-likelihood analysis. FSTR CL 2 consisted of 522 accessions, with highest representation from the Tropical *japonica* (310) and *japonica* groups (94); its accessions were similar to those of CL IIa from the maximum-likelihood analysis. Some *indica* (47), Temperate *japonica* (35), and Intermediate (27) accessions were also found in FSTR CL 2. The smallest cluster was FSTR CL 3, whose 90 accessions were dominated by the Aromatic group (50), followed by the Intermediate group (19); it showed congruence with CL Iic from the maximum-likelihood analysis. The largest cluster was FSTR CL 4, which contained 973 accessions and was dominated by the *indica* group (885), with minor contributions from the Tropical *japonica* (26), Intermediate (21), Temperate *japonica* (17), and *aus/boro* (14) groups. FSTR CL 5 contained 372 accessions and was dominated by the Temperate *japonica* group (248), with minor contributions from the Tropical *japonica* (35), Intermediate (29), *indica* (26), and *japonica* (25) groups; it was similar to CL Ib from the maximum-likelihood analysis. FSTR CL 6 consisted of 323 accessions and was dominated by *indica* varieties (297), and FSTR CL 7 consisted of 505 accessions with major contributions from *indica* (451) and Intermediate varieties (27). FSTR CLs 4, 6, and 7 together corresponded to CL Ia from the maximum-likelihood dendrogram, suggesting that the CL Ia accessions could be further divided into three subgroups. The numbers of accessions that constituted different clusters in the FastSTRUCTURE analysis are provided in Supplemental Table 7.

Next, we looked for admixed genotypes in all groups and found that 41% (1242) of the accessions were admixed in nature (Supplemental Table 8). Among all the clusters, the 505 accessions of FSTR CL 7 contained more admixed individuals (351) than pure individuals (154 accessions), followed by FSTR CL 6 (145 admixed and 148 pure accessions; Supplemental Table 8). Assessment of admixtures within varietal groups revealed that the Intermediate category contained more admixtures (94) than pure (41) accessions, whereas the *indica* population had 821 admixed individuals (47%) out of 1743 accessions. Analysis of regional gene pools revealed that only the European region had more admixed (65) than pure individuals (53) (Supplemental Table 8). Distribution of the 520 CC3 accessions in different clusters of the FastSTRUCTURE analysis (FSTR CL 1–7) was assessed to determine the representation of individuals from each cluster in the mini-core collection. CC3 captured 50 (40 pure individuals with Q value >80%) of the 219 FSTR CL 1 accessions (Supplemental Table 9), 42 (23 pure individuals) of the 522 FSTR CL 2 accessions, 24 (13 pure...
Design of a Rice Mini-Core for Association Studies

To avoid spurious marker–trait associations, an association panel should contain nucleotide diversity ($\pi$) equivalent to that of a larger panel, as well as low population structure and low kinship among its members (Yu and Buckler, 2006; Zhu et al., 2008; Yang et al., 2010; Nachimuthu et al., 2015). We therefore performed nucleotide diversity, population structure, and kinship analyses for the mini-core CC3 collection to assess its utility as an association panel.

Nucleotide Diversity and Population Structure Analysis of Mini-Core CC3

To determine whether mini-core C3 recapitulated the nucleotide diversity of the original collection, we tested important genes known to be associated with various traits, including GW, cooking quality, grain color, grain size, flowering time, and panicle development. We found that the nucleotide diversity within these important genes was comparable in both panels (Supplemental Table 10). This result suggests that mini-core CC3, despite being a smaller subset, captures the essential nucleotide diversity of the larger panel. Next, the underlying population structure of the 520 mini-core CC3 accessions was estimated using FastSTRUCTURE, which grouped them into seven clusters ($K = 7$) named CC CL1–CC CL7 (Figure 4B and Supplemental Table 11). CC CL1 contained 73 accessions and was dominated by the indica (57) and Intermediate (12) groups. CC CL2 was the smallest of the clusters; it contained 23 accessions and was dominated by the Aromatic (9) and Intermediate (8) groups. CC CL3 contained 49 accessions and was dominated by aus/boro (41). CC CL4 contained 70 accessions and had representatives from the indica (58 accesses) and Intermediate (8) groups. CC CL5 contained 43 accesses from different varietal groups, including the Tropical japonica (17), japonica (10), and Intermediate (8) groups. CC CL6 was the largest cluster and contained 191 accesses predominantly from the indica (165) and Intermediate (11) groups. CC CL7 contained 77 accesses mainly from the Temperate japonica (36), Intermediate (14), and japonica (9) groups. A detailed distribution of the accesses from different varietal groups in the seven clusters of mini-core collection CC3 is presented in Supplemental Table 11.

We found that 47% (245) of the accesses in the CC3 mini-core collection were admixtures (Supplemental Table 12). CC CL1 contained 73 accessions and had more admixed individuals (48) than pure individuals (25), followed by CC CL4 with 70 accessions (42 admixed and 28 pure individuals) (Supplemental Table 12). Clusters CC CL2, 5, and 7 had approximately equal numbers of pure and admixed accesses, whereas CC CL3 and CC CL6 had more pure individuals than admixed individuals. Admixture assessment of the varietal groups in CC3 revealed that the Intermediate group had more admixtures (47) than pure (14) individuals, followed by japonica with 12 admixed and 11 pure individuals. The indica group had 144 admixed genotypes (49%) out of 295 accesses, and the Tropical japonica group had 13 admixed accesses out of 27 individuals. Analysis of rice accesses from different regional gene pools in CC3 revealed that South East Asia (63 admixed and 60 pure accesses), China (54 admixed and 47 pure accesses), America (23 admixed and 15 pure accesses), Europe (13 admixed and 6 pure accesses), and Oceania (3 admixed and 2 pure accesses) gene pools contained more admixed individuals than pure individuals (Supplemental Table 12). A detailed distribution of admixed and pure accesses from the different groups present in CC3 is provided in Supplemental Table 12. An increased number of admixed individuals in the clusters derived from population structuring (CC CL1–CC CL7), varietal, and geographic identities confirms that CC3 contained more unrelated individuals than the original collection and validates its suitability as an association panel.

Kinship Analysis of Mini-Core CC3 Individuals

Kinship analysis between individuals from mini-core collection CC3 was performed to estimate their co-ancestry. Seventy percent of the possible pairs of CC3 accessions had kinship...
values less than zero, whereas 25.4% of the accession pairs had kinship values ranging between 0% and 0.25% (Figure 5).

Approximately 4.5% of the CC3 accession pairs showed kinship values in the range of 0.25%–0.50%, and only 0.1% of accession pairs had kinship values in the range of 0.5%–0.75% (Figure 5). Thus, the kinship values for most CC3 accessions exhibited an absence or a weak level of genetic relatedness, fulfilling the primary requirement for utilization of the CC3 mini-core collection as an association panel.

GWAS of Mini-Core Collection CC3

Because mini-core collection CC3 showed low population structure and low kinship values, we proceeded to study its utility for GWAS in rice. GWAS was performed on 520 CC3 mini-core accessions using a compressed mixed linear model (MLM) with 2,081,521 SNPs (MAF >0.02) and 18 yield-related traits of agronomic importance. Association between markers and traits was considered to be significant at $P < 1 \times 10^{-8}$, with a false discovery rate (FDR) adjusted $P$ value of $2.8 \times 10^{-4}$, and explained 34.2% of the phenotypic variation (Figure 6B and Table 1). This SNP was associated with the gene gSW5/GWS5, which has a well-established correlation with GW (Shomura et al., 2008). ET showed significant associations with 3651 SNPs on chromosomes 2, 4, 6, 8, 11, and 12. The most significant SNP (G/T) was located on chromosome 6 at position 6,294,468, had an FDR-adjusted $P$ value of $1.2 \times 10^{-8}$, and explained 29% of the phenotypic variation (Figure 6D and Table 1). The other significant SNP (T/G) associated with ET was located on chromosome 2 at position 7,531,549, had an FDR-adjusted $P$ value of $8.8 \times 10^{-8}$, and explained 30.9% of the phenotypic variation. Another SNP showing significant association with ET was located on chromosome 5 at position 5,646,385 (C/G), had an FDR-adjusted $P$ value of $1.6 \times 10^{-8}$, and explained 55.1% of the phenotypic variation (Table 1). SCC was associated with 306 SNPs on chromosome 7. The most significant SNP (C/G) associated with SCC was located at position 6,526,385, had an FDR-adjusted $P$ value of $4.5 \times 10^{-8}$, and explained 34.2% of the phenotypic variation (Figure 6B and Table 1). This SNP was associated with the $Rc$ gene described in a previous report as an important locus for SCC (Sweeney et al., 2006). Another SNP (T/G) significantly associated with SCC was located at position 6,660,825 on chromosome 7, had an FDR-adjusted $P$ value of $1.6 \times 10^{-6}$, and explained 34.2% of the phenotypic variation (Table 1). Two SNPs significantly associated with SB were identified on chromosome 4 at

![Figure 5. Kinship Analysis of the 520 Accessions from Mini-Core CC3.](image-url)
Table 1. List of SNPs that Showed Significant Associations with Different Traits Identified in Mini-Core Collection CC3 and in the Original Collection of 3004 Rice Accessions.

Nipp, Nipponbare; Chr, chromosome.
* and # represent the mini-core and original collection association markers, respectively. $ represents the association markers found exclusively in the mini-core subset CC3.

| Trait              | Chr | Position   | Major allele | Minor allele | Minor allele frequency | Nipp. allele | FDR-adjusted P value | R² value (%) | Known loci |
|--------------------|-----|------------|--------------|--------------|------------------------|--------------|-----------------------|--------------|------------|
| *Grain length      | 3   | 16 733 441 | G            | T            | 0.36                   | G            | 1.4 × 10^-3           | 32.3         | GS3        |
| *Grain length      | 3   | 16 733 441 | G            | T            | 0.36                   | G            | 3.4 × 10^-43          | 43.5         | GS3        |
| *Grain width       | 5   | 5 371 949  | C            | G            | 0.46                   | C            | 2.8 × 10^-4           | 34.2         | qSW5       |
| *Grain width       | 5   | 5 371 866  | C            | T            | 0.49                   | C            | 9.3 × 10^-34          | 51.4         | qSW5       |
| *Endosperm type    | 6   | 1 765 761  | T            | G            | 0.13                   | T            | 6.4 × 10^-8           | 25           | Waxy       |
| *Endosperm type    | 6   | 1 731 808  | G            | C            | 0.20                   | T            | 1.03 × 10^-29         | 20.2         | Waxy       |
| *Endosperm type    | 6   | 6 294 468  | G            | T            | 0.07                   | G            | 1.2 × 10^-8           | 29           |           |
| *Endosperm type    | 6   | 6 830 286  | G            | A            | 0.21                   | G            | 3.4 × 10^-8           | 15.6         |           |
| *Endosperm type    | 6   | 7 413 964  | C            | G            | 0.24                   | C            | 8.8 × 10^-6           | 20.3         |           |
| *Endosperm type    | 7   | 6 124 457  | C            | T            | 0.456                  | T            | 4.5 × 10^-8           | 61.6         | Rc         |
| *Seed coat color   | 6   | 6 606 825  | T            | G            | 0.454                  | T            | 1.6 × 10^-6           | 59.7         |           |
| *Seed coat color   | 7   | 6 656 052  | T            | G            | 0.43                   | T            | 1.8 × 10^-8           | 6.8          |           |
| *Secondary branching | 2   | 5 032 535 | C            | T            | 0.013                  | C            | 6.4 × 10^-7           | 32           |           |
| *Secondary branching | 4   | 2 521 459 | A            | G            | 0.052                  | A            | 1.6 × 10^-4           | 23.5         |           |
| *Secondary branching | 4   | 12 427 420 | G            | A            | 0.208                  | G            | 1.6 × 10^-4           | 23.4         |           |
| *Panicle axis      | 4   | 1 075 655  | A            | C            | 0.013                  | A            | 3.7 × 10^-4           | 24           |           |
| *Panicle axis      | 6   | 28 676 456 | G            | A            | 0.0078                 | G            | 3.7 × 10^-4           | 23.1         |           |
| *Panicle axis      | 10  | 14 829 875 | C            | A            | 0.0078                 | C            | 3.7 × 10^-4           | 23           |           |

Next, we evaluated the utility of the mini-core collection for a trait other than yield. Fortunately, salt-tolerance data are now available for the original panel accessions. We therefore performed GWAS for salt injury (EC18) using the 520 mini-core CC3 accessions and identified seven SNPs that showed significant association with salt tolerance. These SNPs were distributed on chromosomes 1, 5, 6, 8, 9, 11, and 12. Of the seven SNPs, five had previously reported associations with salt-tolerance- or abiotic-stress-linked QTLs (saltol, qSMC1, qCILV-8.1a, qSISIFH-8.1, qSSISYG5.1, and qSSISYG6.2) (Pandit et al., 2010; Tiwari et al., 2016; Naveed et al., 2018). The details of the SNPs associated with salt stress are

GWAS Using the Original Panel of 3004 Accessions
To further validate the efficiency of the CC3 mini-core collection in capturing the maximum number of marker–trait associations, we performed GWAS analyses for the same yield traits using the original collection of 3004 rice accessions covering genome-wide SNPs. The number of SNPs was reduced due to limitations on matrix size in the R program. For the original collection, 1790 SNPs were significantly associated with different traits

positions 2 521 459 (A/G) and 12 427 420 (G/A). They had an FDR-adjusted P value of 1.6 × 10^-4 and explained 23.5% and 23.4% of the phenotypic variation, respectively. PA was associated with 121 SNPs on chromosomes 2, 4, 6, and 10. The most significant SNP (A/C) was located on chromosome 4 at position 1 075 655, had an FDR-adjusted P value of 3.7 × 10^-4, and explained 24% of the phenotypic variation (Supplemental Figure 1 and Table 1). SNPs on chromosomes 6 and 10 at positions 28 676 456 (G/A) and 14 829 875 (C/A) also showed an association with PA. They had an FDR-adjusted P value of 3.7 × 10^-4 and explained 23.1% and 23.4% of the phenotypic variation, respectively.

Next, we evaluated the utility of the mini-core collection for a trait other than yield. Fortunately, salt-tolerance data are now available for the original panel accessions. We therefore performed GWAS for salt injury (EC18) using the 520 mini-core CC3 accessions and identified seven SNPs that showed significant association with salt tolerance. These SNPs were distributed on chromosomes 1, 5, 6, 8, 9, 11, and 12. Of the seven SNPs, five had previously reported associations with salt-tolerance- or abiotic-stress-linked QTLs (saltol, qSMC1, qCILV-8.1a, qSISIFH-8.1, qSSISYG5.1, and qSSISYG6.2) (Pandit et al., 2010; Tiwari et al., 2016; Naveed et al., 2018). The details of the SNPs associated with salt stress are

Two recent studies evaluated the salinity tolerance of 191 Temperate japonica accessions from the 3KRG panel and identified one overlapping QTL, qPD18_11 and qSES18_11 (Batayeva et al., 2018). There were 24 accessions in common between this panel of 191 Temperate japonica accessions and mini-core CC3 designed in the present study (Supplemental Table 14). This result suggests that mini-core CC3 is also suitable for studying other traits. The identification of previously characterized QTLs for yield traits confirms the utility and importance of mini-core CC3. A detailed description of the SNPs that showed significant associations with ET, GL, GW, PA, SB, and SCC in the CC3 GWAS analysis is presented in Table 1.
Figure 6. Genome-wide Mapping of SNPs Associated with Different Yield-Related Traits in Accessions from Mini-Core CC3.

Manhattan (left) and Q–Q (right) plots of compressed MLM for (A) grain length, (B) grain width, (C) endosperm type, (D) seed coat color, and (E) secondary branching. Negative log10-transformed $P$ values ($y$ axis) from the compressed MLM are plotted against the positions of SNPs ($x$ axis) on different chromosomes. The green line in each figure represents the genome-wide cutoff for significant association. Red and blue lines in the Q–Q plot represent the trajectory for the null hypothesis and the observed values, respectively.
adjusted

and explained from 5.6% to 51.4% of phenotypic variation. Notably, four of the six traits (ET, GL, GW, and SCC) showed significant marker–trait association. However, two traits, 100 grain weight (HGW) and panicle threshability (PT), showed associations in the analysis of the original collection but were missing in the analysis of the CC3 mini-core (Supplemental Table 15). GL was associated with 325 SNPs on chromosomes 3 and 5. The most significant SNP (G/T) associated with GL was located on chromosome 3 at position 16 733 441, had an FDR-adjusted P value of 3.4 × 10⁻⁴, and explained 43.5% of the phenotypic variation (Supplemental Figure 2A and Supplemental Table 15). Another SNP associated with GL was located on chromosome 5 (G/A) at position 5 361 894, had an FDR-adjusted P value of 1.03 × 10⁻³, and explained 38.8% of the phenotypic variation. GW was associated with 737 SNPs on chromosome 5. Consistent with the earlier studies, the most significant SNP on chromosome 5 (G/A) at position 5 371 686, had an FDR-adjusted P value of 9.3 × 10⁻³⁴, and explained 51.4% of the phenotypic variation (Supplemental Figure 2B and Supplemental Table 15). The second SNP (T/C) associated with GW on chromosome 5 was located at position 5 381 894, had an FDR-adjusted P value of 7.9 × 10⁻⁵, and explained 35.2% of the phenotypic variation. Another SNP (T/C) was present on chromosome 5 at position 5 375 201, had an FDR-adjusted P value of 7.9 × 10⁻⁵, and explained 35.2% of the phenotypic variation (Supplemental Figure 2C and Supplemental Table 15). ET was associated with 503 SNPs on chromosome 6. The most significant SNP (G/T) was located on chromosome 6 at position 1 731 808, had an FDR-adjusted P value of 1.03 × 10⁻²⁹, and explained 20.2% of the phenotypic variation (Supplemental Figure 2D). The next significant SNP (G/A) was identified at position 6 830 286, had an FDR-adjusted P value of 3.4 × 10⁻¹⁰, and explained 15.6% of the phenotypic variation. Several SNPs significantly associated with SCC were identified on chromosomes 2 and 7. The most significant SNP was located on chromosome 7 (G/A) at position 6 133 394, had an FDR-adjusted P value of 6.6 × 10⁻¹¹, and explained 7.2% of the phenotypic variation (Supplemental Figure 2E). Three additional SNPs were associated with SCC: SNP (G/T, 6 417 000) and SNP (T/C, 6 656 052) on chromosome 7 and SNP (A/G, 32 431 463) on chromosome 2. These three associations explained between 5.6% and 7.1% of the phenotypic variation. One SNP (C/T) on chromosome 2 showed a significant association with PT; it had an FDR-adjusted P value of 6.8 × 10⁻³ and explained 16.4% of the phenotypic variation (Supplemental Figure 2F). A detailed distribution of the SNPs associated with traits such as ET, GL, GW, HGW, PT, and SCC in the original panel of 3004 accessions is provided in Supplemental Table 15.

**Linkage Disequilibrium and Haplotype Analysis**

To gain further insight into some of the less well-characterized marker–trait associations identified in the CC3 panel, we studied the associated SNPs for LD pattern and haplotype block formation using their flanking nucleotides. In the case of trait SB, 100 SNPs flanking the most significantly associated SNP (A/G, 2 521 459) on chromosome 4 were used, and an LD block containing the associated SNP was identified. The formulated block showed strong LD in a span of 2 kb that contained seven neighboring SNPs, including the A/G at 2 521 459 (Figure 7A). Haplotype analysis of this block revealed that the PSB_H1 (ATCAGGT) haplotype had the highest frequency (f = 0.45). Distribution of haplotypes between light and dense panicle secondary branching revealed that all but the H4 haplotype had significant associations with light-level branching (Figure 7A). By contrast, the H4 haplotype showed an inclination toward a dense branching trait. One recent study has demonstrated the association of elevated haplotype diversity in SHORT PANICLE 1 (SP1) with phenotype in the japonica rice group (Jang et al., 2018). Similarly, a GL-associated SNP (G/T, 16 733 441) on chromosome 3 was also analyzed for haplotype mining. LD analysis was performed to identify a block that contained the associated SNP. This block showed strong LD within a span of 1 kb that contained three neighboring SNPs, including the associated one (Figure 7B). Haplotype analysis of this block revealed that the GL_H1 (TTG) haplotype had the highest frequency (f = 0.631) of all the haplotypes (GL_H2; TTT = 0.315, GL_H3; TCG = 0.034, GL_H4; CCG = 0.013). Distribution of these haplotypes between long- and short-grain accessions revealed that GL_H2, in addition to being the second most frequent haplotype, was also maximally associated with long grains, with an average of 9.33 mm GL. Similarly, haplotypes GL_H1 and GL_H3 were linked to intermediate GL, with mean values of 8.2 and 7.9 mm, respectively (Figure 7B). On the other hand, the H4 haplotype showed an inclination toward the short-grain trait, with the lowest mean GL of 4.8 mm. This haplotyping observation was similar to that of a previous study that dissected the separate clustering of grain-length haplotypes for varying size and different rice groups (Singh et al., 2017).

**Concluding Remarks**

Despite tremendous efforts, the resolution of QTLs responsible for yield-related traits and their causative genes has remained limited due to their complex, multifactorial nature. QTL mapping using a diverse panel and GWAS analysis have been proven to be effective tools for understanding the genetic basis of any trait. For GWAS analysis, estimation of the underlying population structure of the panel under consideration is important and helps to avoid spurious associations between phenotypes and genotypes (Pritchard and Rosenberg, 1999; Pritchard et al., 2000; Pritchard and Donnelly, 2001). Most earlier studies in rice have considered a particular population (Huang et al., 2010; Lu et al., 2015) that may have had a high level of structure and kinship affecting the GWAS analysis and resulting in spurious marker–trait associations. This study represents the first time that a large set of diversified rice germplasms (3004) was used, providing complete coverage of the global rice gene pool. The mini-core was developed using more than 2 million genome-wide SNPs, 18 different phenotypes, and 89 country locations. The mini-core accounted for 17.3% of the original collection and captured the maximum SNP polymorphism. All the original phenotypes and geographic regions were represented in the mini-core. The mini-core showed nucleotide diversity equivalent to that of the original panel, as well as low population structure and low or no kinship among individuals, thereby...
avoiding spurious marker–trait associations. Furthermore, an increase in the number of admixed individuals in different clusters of the CC3 structure analysis showed that the panel was unstructured and diverse in nature, appropriate for use in association analysis.

On the utility front, GWAS with the mini-core panel identified various novel marker–trait associations and validated earlier reported associations. This analysis also provided a tool for comparison between the CC3 mini-core and the original collection. We were able to show that CC3 captured the associations prevalent in the original collection and was therefore a representative subset. In conclusion, we were able to generate and validate mini-core CC3 as a robust, diversified, non-redundant, and manageable association panel that efficiently mirrored the large collection of 3004 diverse rice accessions. We suggest that this relatively small subset can be used effectively for efficient agronomic trait evaluation, which in turn will be useful for marker-assisted breeding programs for rice crop improvement.

METHODS

Genotypic and Phenotypic Data of the Rice Germplasm Collection

We used SNP data from 3004 rice accessions (hereafter referred to as the original collection) and phenotypic data for 18 yield-related traits (DEH, HGW, ET, DFF, GL, GW, leaf senescence, PA, PL, panicle shattering, PT, SB, SCC, SH, spikelet fertility, culm length, culm number, and culm diameter) to develop mini-core collections and perform association analyses. The 3000 Rice Genome Project (3K RGP) data for 18.9 million polymorphic SNPs and associated phenotypic data were retrieved from the SNP-Seek database (http://snp-seek.irri.org) (Alexandrov et al., 2015; Mansueto et al., 2017). In addition, we performed whole-genome sequencing of four Indian accessions (LGR, PB-1121, Sonasal, and Bindli) at a depth of 45X and collected phenotypic data for the aforementioned traits during the 2016 and 2017 growing seasons. The original collection of rice accessions came from 89 countries and represented all the regional pools and varieties of rice grown throughout the world.

Isolation of Genomic DNA, Genome Sequencing, and SNP Calling

The four Indian rice accessions (long grain: LGR [LG] and PB 1121 [PB]; short grain: Sonasal [SN] and Bindli [BN]) were grown in a research field at the National Institute of Plant Genome Research in 2016. Ten-day-old rice seedlings were used for the isolation of genomic DNA with the Sigma GenElute Plant genomic DNA kit. The integrity of the genomic DNA was analyzed using a 2100 Bioanalyzer (Agilent Technologies, Singapore). Samples for sequencing were prepared using the Illumina TruSeq DNA sample preparation kit (Illumina, USA). Sequencing was performed with 90-bp paired-end chemistry on an Illumina HiSeq 2000 instrument.

Raw reads were quality-checked, and low-quality bases (Phred score < Q30) were removed. The filtered reads were then mapped to the rice Nipponbare reference genome (IRGSP-1.0 pseudomolecule/MSU7) using the BWA program with the –q20 setting. The Picard program was used to remove duplicate reads.

Variant calling of SNPs and InDels was performed using the Genome Analysis TKLite-2.3-9 Unified Genotyper (GATK) (McKenna et al., 2010). SNPs and InDels with a polymorphism call rate of <90% were eliminated. After calling, total variants were stringently filtered based on a read depth threshold of >10X. The filtered variants were then annotated using the SnpEff tool (Cingolani et al., 2012) and the Rice Annotation Project database (http://rapdb.dna.affrc.go.jp).

HGW, ET, DFF, GL, GW, leaf senescence, PA, PL, panicle shattering, PT, SB, SCC, SH, spikelet fertility, culm length, culm number, and culm diameter) to develop mini-core collections and perform association analyses. The 3000 Rice Genome Project (3K RGP) data for 18.9 million polymorphic SNPs and associated phenotypic data were retrieved from the SNP-Seek database (http://snp-seek.irri.org) (Alexandrov et al., 2015; Mansueto et al., 2017). In addition, we performed whole-genome sequencing of four Indian accessions (LGR, PB-1121, Sonasal, and Bindli) at a depth of 45X and collected phenotypic data for the aforementioned traits during the 2016 and 2017 growing seasons. The original collection of rice accessions came from 89 countries and represented all the regional pools and varieties of rice grown throughout the world.

Isolation of Genomic DNA, Genome Sequencing, and SNP Calling

The four Indian rice accessions (long grain: LGR [LG] and PB 1121 [PB]; short grain: Sonasal [SN] and Bindli [BN]) were grown in a research field at the National Institute of Plant Genome Research in 2016. Ten-day-old rice seedlings were used for the isolation of genomic DNA with the Sigma GenElute Plant genomic DNA kit. The integrity of the genomic DNA was analyzed using a 2100 Bioanalyzer (Agilent Technologies, Singapore). Samples for sequencing were prepared using the Illumina TruSeq DNA sample preparation kit (Illumina, USA). Sequencing was performed with 90-bp paired-end chemistry on an Illumina HiSeq 2000 instrument.

Raw reads were quality-checked, and low-quality bases (Phred score < Q30) were removed. The filtered reads were then mapped to the rice Nipponbare reference genome (IRGSP-1.0 pseudomolecule/MSU7) using the BWA program with the –q20 setting. The Picard program was used to remove duplicate reads.

Variant calling of SNPs and InDels was performed using the Genome Analysis TKLite-2.3-9 Unified Genotyper (GATK) (McKenna et al., 2010). SNPs and InDels with a polymorphism call rate of <90% were eliminated. After calling, total variants were stringently filtered based on a read depth threshold of >10X. The filtered variants were then annotated using the SnpEff tool (Cingolani et al., 2012) and the Rice Annotation Project database (http://rapdb.dna.affrc.go.jp).
Four-gamete-rule method was employed to identify the more refined genomic block that contained the associated SNPs.

**SUPPLEMENTAL INFORMATION**

Supplemental Information can be found online at *Plant Communications Online.*

**ACKNOWLEDGMENTS**

A.K. acknowledges the University Grant Commission, Government of India and NIPGR for the Research Fellowships. S.K. acknowledges a National Postdoctoral Fellowship from the Science and Engineering Research Board, Department of Science and Technology, Government of India and a Short-Term Research Fellowship from NIPGR. K.B.M.S. acknowledges the Council of Scientific and Industrial Research, Government of India for the Junior Research Fellowship. The authors are grateful to the DBT-eLibrary Consortium for providing access to literature. No conflict of interest declared.

Received: October 31, 2019

Revised: December 13, 2019

Accepted: April 21, 2020

Published: April 24, 2020

**REFERENCES**

Agrawal, H.A., Yan, W., Lee, F., Fjellstrom, R., Chen, M.-H., Jia, M., and McClung, A. (2009). Genetic assessment of a mini-core subset developed from the USDA rice genebank. Crop Sci. 49:1336.

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R.R., Ulat, V.J., Chebotarov, D., Zhang, G., Li, Z., et al. (2015). SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 43:D1023–D1027.

Ambreen, H., Kumar, S., Kumar, A., Agarwal, M., Jagannath, A., and Goel, S. (2018). Association mapping for important agronomic traits in safflower (*Carthamus tinctorius L*) core collection using microsatellite markers. Front. Plant Sci. 9:402.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Batayeva, D., Labaco, B., Ye, C., Li, X., Usenbekov, B., Rysbekova, A., Dyuskalieva, G., Vergara, G., Reinke, R., and Leung, H. (2018). Genome-wide association study of seedling stage salinity tolerance in temperate japonica rice germplasm. BMC Genet. https://doi.org/10.1186/s12863-017-0590-7.

Bresheghello, F., and Sorrells, M.E. (2006). Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci. 46:1323.
Design of a Rice Mini-Core for Association Studies

Brown, A.H.D. (2011). Core collections: a practical approach to genetic resources management. Genome 31:818–824.

Brown, A.H.D., and Spillane, C. (1999). Implementing core collections — principles, procedures, progress, problems and promise. In Core Collections for Today and Tomorrow, R.C. Johnson and T. Hodgkin, eds. (Rome: IPGRI), pp. 7–17.

De Beukelaer, H., Davenport, G.F., and Fack, V. (2018). Core Hunter 3: flexible core subset selection. BMC Bioinformatics 19:203.

Eban, K., Kojima, Y., Fukuoka, S., Nagamine, T., and Kawase, M. (2008). Development of mini core collection of Japanese rice landrace. Breed. Sci 58:281–291.

Edgerton, M.D. (2009). Increasing crop productivity to meet global needs for feed, food, and fuel. Plant Physiol. 149:7–13.

Eizenga, G.C., Ali, M.L., Bryant, R.J., Yeater, K.M., McClung, A.M., and McCouch, S.R. (2014). Registration of the rice diversity panel 1 for genomewide association studies. J. Plant Registr. 8:109.

Eizenga, G.C., Ali, M.L., Bryant, R.J., Yeater, K.M., McClung, A.M., and McCouch, S.R. (2014). Registration of the Rice Diversity Panel 1 for Genomewide Association Studies. J. Plant Registr. 8 (1):109–116.

El Bakkali, A., Haouane, H., Moukhli, A., Costes, E., Van Damme, P., and Khadari, B. (2013). Construction of core collections suitable for association mapping to optimize use of mediterranean olive (Olea europaea L.) genetic resources. PLoS One 8:e61265.

Famoso, A.N., Zhao, K., Clark, R.T., Tung, C.-W., Wright, M.H., Bustamante, C., Kochian, L.V., and McCouch, S.R. (2011). Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genomewide association analysis and QTL mapping. Plos Genet. 7:e1002221.

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., and Zhang, Q. (2006). GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor. Appl. Genet. 112:1164–1171.

Food and Agriculture Organization of the United Nations. FAOSTAT Database. 2017.

Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., et al. (2019). Structural variants in 3000 rice genomes. Genome Res. 29:870–880.

Gao, Z. (2003). Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. Sci. China Ser. C 46:661.

Gupta, P.K., Rustgi, S., and Kulwal, P.L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol. Biol. 57:461–485.

Hammer, Ø., Harper, D., and Ryan, P. (2001). Past: paleontological statistics software package for education and data analysis. Paleontol. Electron. 4. http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

Hardy, O.J., and Vekemans, X. (2002). spagetti: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

Hu, J., Zhu, J., and Xu, H.M. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor. Appl. Genet. 101 (1–2):264–268.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., and et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., and et al. (2012). Genome-wide association analysis of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat. Genet. 44:32–39.

Ingvarsson, P.K., and Street, N.R. (2011). Association genetics of complex traits in plants. New Phytol. 189:909–922.

Jang, S., Lee, Y., Lee, G., Seo, J., Lee, D., Yu, Y., Chin, J.H., and Koh, H.-J. (2018). Association between sequence variants in panicle development genes and the number of spikelets per panicle in rice. BMC Genet. 19:5.

Korte, A., Vilhjalmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. 44:1066–1071.

Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M.A., Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat. Genet. 43:163–168.

Lee, T.-H., Guo, H., Wang, X., Kim, C., and Paterson, A.H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162.

Li, Z., Rutger, J., Yu, S., Xu, W., Vijayakumar, C., Ali, J., Fu, B., Xu, J., Marghirang, R., Domingo, J., et al. (2014). The 3,000 rice genomes project. Gigascience 3:7.

Liakat Ali, M., McCouch, A.M., Jia, M.H., Kimball, J.A., McCouch, S.R., and Eizenga, G.C. (2011). A rice diversity panel evaluated for genetic and agro-morphological diversity between subpopulations and its geographic distribution. Crop Sci. https://doi.org/10.2135/cropsci2010.11.0641.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399.

Lu, Q., Zhang, M., Niu, X., Wang, S., Xu, Q., Feng, Y., Wang, C., Deng, H., Yuan, X., Yu, H., et al. (2015). Genetic variation and association mapping for 12 agronomic traits in indica rice. BMC Genomics 16:1067.

Mansueto, L., Fuentes, R.R., Chebotarov, D., Borja, F.N., Detras, J., Abriol-Santos, J.M., Palis, K., Poliakov, A., Dubchak, I., Solovyev, V., et al. (2016). SNP-Seek II: a resource for allele mining and analysis of big genomic data in Oryza sativa. Curr. Plant Biol. 7:8–16.

Mansueto, L., Fuentes, R.R., Borja, F.N., Detras, J., Abriol-Santos, J.M., Chebotarov, D., Sanciangco, M., Palis, K., Copetti, D., Poliakov, A., et al. (2017). Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. 45:D1075–D1081.

McCouch, S.R., Wright, M.H., Tung, C.-W., Marong, L.M., McNally, K.L., Fitzgerald, M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., et al. (2016). Open access resources for genome-wide association mapping in rice. Nat. Commun. 7:10532.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McNally, K.L., Childs, K.L., Bohnert, R., Davidson, R.M., Zhao, K., Ulat, V.J., Zeller, G., Clark, R.M., Hoen, D.R., Bureau, T.E., et al. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc. Natl. Acad. Sci. U S A 106:12273–12278.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2012). Crop genomics: advances and applications. Nat. Rev. Genet. 13:85–96.

Nachimuthu, V.V., Muthurajan, R., Duraialaguraja, S., Sivakami, R., Pandian, B.A., Ponniah, G., Gunasekaran, K., Swaminathan, M., Suji, K., and Sabariappan, R. (2015). Analysis of population
structure and genetic diversity in rice germplasm using SSR markers: an initiative towards association mapping of agronomic traits in *Oryza sativa*. Rice 8:30.

Naveed, S.A., Zhang, F., Zhang, J., Zheng, T.Q., Meng, L.J., Pang, Y.L., Xu, J.L., and Li, Z.K. (2018). Identification of QTN and candidate genes for salinity tolerance at the germination and seedling stages in rice by genome-wide association analyses. Sci. Rep. 8:6505.

Pandit, A., Rai, V., Bal, S., Sinha, S., Kumar, V., Chauhan, M., Gautam, R.K., Singh, R., Sharma, P.C., Singh, A.K., et al. (2010). Combining QTL mapping and transcriptome profiling of bulked RILs for identification of functional polymorphism for salt tolerance genes in rice (*Oryza sativa*). Mol. Genet. Genomics 284:121–136.

Perseguini, J.M.K.C., Silva, G.M.B., Rosa, J.R.B.F., Gazaffi, R., Marçal, J.F., Carbonell, S.A.M., Chiorato, A.F., Zucchi, M.I., Garcia, A.A.F., and Benchimol-Reis, L.L. (2015). Developing a common bean core collection suitable for association mapping studies. Genet. Mol. Biol. 38:67–78.

Pritchard, J.K., and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220–228.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Raj, A., Stephens, M., and Pritchard, J.K. (2014). FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–589.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., and Yano, M. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. Nat. Genet. 40:1029–1028.

Singh, N., Singh, B., Rai, V., Sidhu, S., Singh, A.K., and Singh, N.K. (2017). Evolutionary insights based on SNP haplotypes of red pericarp, grain size and starch synthase genes in wild and cultivated rice. Front. Plant Sci. 8:372.

Sweeney, M.T., Thomson, M.J., Pfeil, B.E., and McCouch, S. (2006). Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell 18:283–294.

Tian, M., Tan, G., Liu, Y., Rong, T., and Huang, Y. (2009). Origin and evolution of Chinese waxy maize: evidence from the Globulin-1 gene. Genet. Resour. Crop Evol. 56:247–255.

Tiwari, S., Sl, K., Kumar, V., Singh, B., Rao, A., Mythra SV, A., Rai, V., Singh, A.K., and Singh, N.K. (2016). Mapping QTLs for salt tolerance in rice (*Oryza sativa* L.) by bulked segregant analysis of recombinant inbred lines using 50K SNP chip. PLoS One 11:e0153610.

Upadhyaya, H.D., Bajaj, D., Das, S., Saxena, M.S., Badoni, S., Kumar, V., Tripathi, S., Gowda, C.L.L., Sharma, S., Tyagi, A.K., et al. (2015). A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. Plant Mol. Biol. 89:403–420.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49.

Yang, X., Yan, J., Shah, T., Warburton, M.L., Li, Q., Li, L., Gao, Y., Chai, Y., Fu, Z., Zhou, Y., et al. (2010). Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor. Appl. Genet. 121:417–431.

Yu, J., and Buckler, E.S. (2006). Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17:155–160.

Zhang, P., Liu, X., Tong, H., Lu, Y., and Li, J. (2014). Association mapping for important agronomic traits in core collection of rice (*Oryza sativa* L.) with SSR markers. PLoS One 9:e111508.

Zhao, K., Tung, C.-W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat. Commun. 2:467.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and prospects of association mapping in plants. Plant Genome J. 1:5.