Whole-Genome Sequencing of the NARO World Rice Core Collection (WRC) as the Basis for Diversity and Association Studies

N. Tanaka¹, M. Shenton¹, Y. Kawahara¹,², M. Kumagai², H. Sakai², H. Kanamori¹, J. Yonemaru¹, S. Fukuoka¹, K. Sugimoto¹, M. Ishimoto¹, J. Wu¹ and K. Ebana³,*

¹Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki 305-8518 Japan
²Advanced Analysis Center, National Agriculture and Food Research Organization, Tsukuba Ibaraki, 305-8517, Japan
³Genetic Resources Center, National Agriculture and Food Research Organization, Plant Genetic Diversity Laboratory, Tsukuba, Ibaraki 305-8502, Japan
*Corresponding author: E-mail, ebana@affrc.go.jp; Fax, +81-29-838-7408.
(Received October 31, 2019; Accepted February 16, 2020)

Introduction

Rice is a staple food on which 50% of the world’s population depends (FAO 2004). The development of new rice cultivars of high yield and quality is crucial for agricultural sustainability and human health (Godfray et al. 2010). Traditionally, biallelic mapping has been used to identify quantitative trait loci (QTLs) related to various agronomic traits in rice (Miura et al. 2011). However, biallelic mapping can only detect the limited genetic diversity that exists in a biparental population. Therefore, it is necessary to exploit more diverse genetic resources to meet the demand for increased environmentally sustainable rice production and to detect the functions of agriculturally important genes derived from those resources. To this end, there has been a trend toward large-scale resequencing of large germplasm collections (Wang et al. 2018), but the use of such large collections requires substantial resources of time and money. Thus, the use of core or mini-core germplasm collections is still vitally important for agronomical research. ‘Core collections’, small sets of germplasm representing the genetic diversity in large genebank collections, are a powerful tool enabling the effective use of germplasm. The core collections of several crops have been chosen on the basis of passport data recorded for each accession, to represent variation in phenotypes and geographical characteristics of the germplasm source (Johnson and Hodgkin 1999).

Previously, the World Rice Core Collection (WRC) was developed in the National Agriculture and Food Research Organization (NARO) Genebank based on the passport data and restriction enzyme fragment length polymorphism (RFLP) markers (Kojima et al. 2005). Because the WRC is a very small population with a high degree of genetic diversity, it has been used for the evaluation of many agricultural traits, such as heading date, seed cadmium concentration and seed dormancy, in >300 research projects (e.g. Takahashi et al. 2009, Ueno et al. 2009, Uraguchi et al. 2009, Sugimoto et al. 2010, Ogiso-Tanaka 2009, Uraguchi et al. 2009, Sugimoto et al. 2010, Ogiso-Tanaka 2009, Uraguchi et al. 2009).
Diversity of the WRC

The NARO WRC, composed of 69 accessions (Supplementary Table S1), is a useful germplasm set for agronomical research, and >300 seed sets have been distributed to date. In this study, we performed WGS of the 69 accessions and obtained detailed nucleotide polymorphism information. The reads were mapped against the Oryza sativa ssp. japonica Nipponbare genome (Os-Nipponbare-Reference-IRGSP-1.0) (Kawahara et al. 2013) using bwa (Li and Durbin 2009). The average depth was 38, ranging from 20 to 102; the average number of SNPs and insertion–deletion (indels) per accession were 1,835,874 and 329,670, respectively (Supplementary Table S1). After removing data for positions where genotype data were missing, and selecting biallelic variants with a minor allele frequency (MAF) of >0.05, a total of 2,805,329 SNPs and 357,639 indels were obtained. Among the variant data, 98,833 variants caused non-synonymous substitutions or otherwise altered the predicted amino acid sequence of protein-coding genes.

We investigated the population structure of the WRC accessions using these genome-wide SNP data. We constructed a phylogenetic tree of the WRC based on a set of 2,315 representative SNPs selected to be in approximate linkage equilibrium (using the SnpPhylo pipeline; Lee et al. 2014), and the WRC accessions were divided into three major groups, corresponding to japonica, indica and aus (Fig. 1A). Among these three groups, the indica cluster could be further divided into four subgroups (I-1, I-2, I-3 and I-4) and the aus cluster could be further divided into two subgroups (A-1 and A-2) based on the phylogenetic tree (Fig. 1A). Principal component analysis (PCA) using the whole SNP dataset also classified the WRC into three groups (Fig. 1B). The contribution of the first and second principal components was 22.6% and 14.2%, respectively. It was inferred that the optimum number of clusters (K) in this population using all SNPs is 3 based on the fastStructure analysis (Fig. 1C, Supplementary Table S1). These results revealed a clear separation of the WRC into three groups consistent with the classification using RFLP markers by Kojima et al. (2005), and thus, the WRC represents a highly structured population. Compared with their RFLP-based classification in Kojima et al. (2005), Lebed (WRC23) (indica in Kojima et al. 2005) was reclassified into japonica and Calotoc (WRC22) and Basilanon (WRC44) (aus in Kojima et al. 2005) were reclassified into the indica subgroup based on the WGS data (Fig. 1A). According to the fastStructure analysis, it is likely that there is some admixture in these accessions and, in the phylogenetic tree shown in Fig. 1A, these accessions are separated from the main branches in japonica and indica subgroups. Therefore, it is not surprising that there was a difference between the classifications based on the WGS and RFLP markers.

To compare the population structure of the WRC with publicly available data from the 3,000 Rice Genomes Project, we downloaded a whole-genome biallelic SNP set from IRRI Snpseek (https://snp-seek.irri.org/) and performed PCA using WRC data for the same SNP positions. A comparison of WRC groups with the rice population groups defined by Wang et al. (2018) suggests that the WRC is an effective mini-core collection for rice, as much of the variation in the first two principal components can be captured by the WRC (Supplementary Fig. S1). A large phylogenetic tree comprising both WRC and 3K accessions can be accessed at https://itol.embl.de/tree/1502676182761580197097. The assignment of WRC accessions to population groups defined by Wang et al. (2018) based on proximity to 3K genome accessions in the phylogenetic tree is shown in Supplementary Table S1. Haplotype analysis of agronomically significant genes

To demonstrate the utility of the WRC genome sequence data, we examined the haplotypes of a series of agronomically important genes. We visualized WGS data of the WRC in the
TASUKE multiple genome browser (Kumagai et al. 2013) and classified haplotype groups of several genes related to seed traits and heading date.

In grain color (Fig. 2B), a major pericarp-color gene Rc (Os07g0211500), which encodes a basic helix–loop–helix protein, has been reported (Furukawa et al. 2006, Sweeney et al. 2006). In 23 of the 69 accessions (12/16 aus, 9/33 indica and 2/20 japonica accessions), we detected a 14-base insertion in exon 6 of Rc (Fig. 2A), known to be a functional mutation (Furukawa et al. 2006) that results in a frameshift mutation, while 20 of these 23 accessions had a nonwhite pericarp. The pericarp color of WRC25 (Muha), WRC26 (Jhona 2) and WRC42 (Local Basmati) was white, even though these accessions carried the 14-base insertion in the Rc gene. All these three accessions, as well as WRC33 (Surjamukhi) that had a light red pericarp, were categorized into aus and also carried a C → A mutation at position Cys 451. This mutation, resulting in a stop codon at amino acid 451 in Rc, is known as the Rc-s allele (Sweeney et al. 2007). The Rc-s allele is known to be specific to aus and results in white or light red pericarps (Sweeney et al. 2007), consistent with our results.

As for amylose content, the waxy (Wx) gene (Os06g0133000) encodes a granule-bound starch synthase and controls the synthesis of amylose in endosperm (Wang et al. 1995). We found a 23-bp insertion in exon 1 of the waxy gene, resulting in a frameshift in nine accessions with low amylose content (3/33 indica and 6/20 japonica accessions; Fig. 2C, D). This 23-bp insertion in waxy was completely correlated with the amylose content in WRC seeds (Fig. 2D) and has been described previously along with several other haplotypes at the waxy locus in wild and cultivated rice (Zhang et al. 2019).

We performed haplotype analysis of GS3 (Os03g0407400), which encodes a protein consisting of four domains and an N-terminal plant-specific organ size regulation (OSR) domain that is sufficient to negatively regulate grain size (Mao et al. 2010). We observed two high-impact mutations in the GS3 sequence, and we first focused on an SNP that changes a Cys codon to a stop codon at amino acid 55 (Fig. 2E). WRC accessions carrying this haplotype B in GS3 had longer grains than haplotype A (Fig. 2F). The other high-impact mutation was seen in WRC34 (ARC 7291), which had a 4-bp deletion in exon 5 of GS3 and showed shorter grain length (5.0 mm, average: 6.9 mm) (Supplementary Table S1). Haplotype analysis for genes related to seed phenotypes indicated that there was functional diversity in well-known genes related to grain color, amylose content and grain length among the 69 WRC accessions.
Heading date is one of the most important agronomical traits, and diversity of heading date has contributed to the development of rice cultivation in a wider range of latitudes (Kush 1997). As major heading date QTLs, Hd1 (Yano et al. 2000), Hd2/OsPRR37 (Koo et al. 2013, Gao et al. 2014), Hd6 (Ogiso et al. 2010), Hd17/OsELF3-1 (Matsubara et al. 2012), RFT1 (Komiya et al. 2008), Ghd7 (Xue et al. 2008) and DTH8 (Wei et al. 2010) have been identified (Itoh et al. 2018). We used the TASUKE system to explore haplotypes of heading date genes in the WRC. Most of the reported alleles were detected in the WRC, with some exceptions. The distribution of alleles generally reflected the origin and/or cultivar groups of the accessions. Among the haplotypes of Hd1, a 2-bp deletion in the hd1 null alleles was widely distributed throughout the japonica, indica and aus groups (Fig. 3, Table 1). This 2-bp deletion, as well as other frameshift mutations, and an SNP that changes an Arg residue to a stop codon at amino acid 358 have been previously identified as functional mutations in Hd1 (Yano et al. 2000, Takahashi et al. 2009). On the other hand, the SNP that changes Arg358 to a stop codon was only distributed in the japonica accessions WRC45 (Ma sho), WRC46 (Khao Nok) and WRC48 (Khau Mac Kho). We also detected a 1-bp deletion with a frameshift in exon 1 of Hd1 in four accessions (Table 1). All these four accessions (WRC20, WRC22, WRC24 and WRC44) originated from the Philippines and were categorized into a small subgroup 1-4 in the indica group (Fig. 1A, Table 1). A 4-bp deletion in the second exon of Hd1 was observed in WRC7 (Davao1), WRC18 (Qingyu), WRC21 (Shwe Nang Gyi), WRC57 (Milyand 23) and WRC98 (Deejaiohaluuo), which belong to the indica group (Fig. 3, Table 1). As minor alleles, only WRC100 (Vandanar) has a 2-bp insertion in exon 2 of Hd1 (Fig. 3, Table 1).

Next, we examined the haplotypes of other major QTLs for heading date, Hd2/OsPRR37, Hd6, Hd17/OsELF3-1, RFT1, Ghd7 and DTH8 in the WRC accessions (Table 1, Supplementary Fig. S2). In Hd2, an 8-bp deletion was observed in the seventh exon of eight accessions classified into the indica group (Table 1, Supplementary Fig. S2). All accessions with this 8-bp deletion in Hd2, except for WRC16, were classified into a small subgroup (I-2) in indica (Fig. 1A, Table 1). WRC2 (Kasalath), WRC31 (Shoni) and WRC38 (ARC 11094) in the aus A-1 subgroup carry an SNP that changes a Gln to a stop codon at amino acid 705. Only WRC66 (Bingala) in indica has a functional variant that changes Tyr to His at amino acid 704 in the Hd2 CCT domain (Koo et al. 2013). Functional mutations in Hd2 were not observed in japonica accessions in the WRC. In Hd6, a nonsynonymous substitution at amino acid 146 resulting in a premature stop codon was only observed in Nipponbare among the WRC population (Table 1). Matsubara et al. (2012) reported that a substitution from Leu to Ser at amino acid 558 of Hd17 was a functional mutation. Except for five accessions, this substitution was observed in all members of the WRC population, but there was no correlation of this allele with the heading date (Table 1, Supplementary Table S1). In RFT1, the I-2 indica accessions have a functional mutation that changes Glu to Lys at amino acid 105 (Table 1). Null mutations in Ghd7 and DTH8 cause extremely early heading date (Itoh et al. 2018). In Ghd7, previously reported functional mutations (Xue et al. 2008) were not detected among WRC accessions, but WRC15 (Co 13) and WRC100 (Vandanar) carried an SNP in the splicing acceptor region (Table 1). WRC3 (Bei Khe), WRC10 (Shuu Sou Shu) and WRC19 (Deng Pao Zhai) have a 1,118-bp deletion (Hori et al. 2015), and WRC43 has a 19-bp deletion in the DTH8 coding sequence, but only WRC10 showed an early-heading date phenotype (Table 1, Supplementary Table S1). In DTH8, WRC17 (Keiboba) and WRC99 (Hong Cheuh Zai) have an 8-bp deletion and WRC7 (Davao1), WRC9 (Ryo Suisan Koumai), WRC21 (Shwe Nang Gyi) and WRC98 (Deejaiohaluuo) have a 1-bp deletion, which causes a frameshift. Three of these five accessions originate from China and show an early heading date phenotype (101–111 d, mean of WRC: 122 d, Supplementary Table S1). These 8- and 1-bp deletions might be new functional mutations in DTH8 that contribute to an early-flowering phenotype in Chinese accessions.

Using the TASUKE system, we detected several high-impact indels and SNPs in heading date-related genes. Almost all mutations in these seven genes related to heading date were observed in particular groups (Table 1). Only a 2-bp deletion
in *Hd1* was widely distributed among the three subgroups in WRC (Supplementary Fig. S1). These results indicate that indels and SNPs with high impact in heading date genes occurred after subgroup differentiation, except for the 2-bp deletion in *Hd1*. In addition, large deletions, such as a 1,118-bp deletion in *DTH8*, can also be detected by the TASUKE system (Table 1). Because genome-wide SNP sets generally focus on biallelic SNPs and ignore long indels, such long deletions and multi-allelic SNPs may be overlooked, but they can be visualized in parallel for different accessions using the TASUKE system. Thus, TASUKE with the WRC accessions is a useful tool for visualizing and analyzing haplotype variation among diverse rice germplasm resources.

**GWAS with the WRC**

One of the most important aims in maintaining genetic resources in genebanks is to use them as materials for crop improvement in breeding programs. To analyze the natural allelic...
variation associated with agronomic traits using diverse populations, GWAS is a powerful and efficient tool. Therefore, we evaluated the WRC as a GWAS population for agriculturally important traits. We tried GWAS for grain (pericarp) color as a qualitative trait, amylose content as a bimodal quantitative trait, grain length as a quantitative trait controlled by a few QTLs and heading date as an example of a quantitative trait controlled by a complicated gene network (Figs. 4, 5).

When considering grain color as a binary phenotype (white or not white), and using two methods, mixed linear model (MLM) and FarmCPU (Liu et al. 2016), a significant peak associated with grain color was detected on the short arm of chromosome 7 using both methods (Fig. 5A, Supplementary Fig. S3). For the grain color phenotype data, we coded the observation data as a numeric value of 1 for white-colored accessions and 2 for nonwhite color accessions including red or purple grains (Fig. 4A). The detected peak was located close to a major pericarp-color gene Rc (Os07g0211500) (Furukawa et al. 2006, Sweeney et al. 2006). These results indicated that our GWAS system with WRC was suitable for detecting genes associated with binary traits, such as grain color.

Next, we performed GWAS for seed amylose content using the WRC population. The distribution of seed amylose content was bimodal (Fig. 4B). We used the same GWAS methods as before, and significant peaks associated with amylose content were detected on chromosome 6 with both methods (Fig. 5B, Supplementary Fig. S4). These peaks corresponded to a known gene responsible for amylose content, waxy (Os06g0133000) (Wang et al. 1995). The identification of the waxy gene indicated that GWAS using WGS of WRC accessions could also be used to identify genes associated with traits showing a bimodal distribution in the study population.

WRC accessions exhibited a normal distribution for grain length (Fig. 4C), and a clear peak for grain length was observed.

**Table 1** Distribution of functional mutations in flowering-time genes in WRC accessions

| QTLs | Indel/SNPs | Substitution | Reference | WRC accessions |
|------|------------|--------------|-----------|----------------|
| Hd1  | 1-bp deletion | Tyr191fs | | WRC20, 22, 24, 44 |
|      | 2-bp deletion | Phe279fs | Yano et al. (2000) | WRC2, 4–6, 11, 13, 17, 23, 25, 28–38, 40–42, 47, 49, 50, 55, 64, 99 |
|      | 2-bp insertion | Asp294fs | | WRC100 |
|      | 4-bp deletion | Lys352fs | | WRC7, 18, 21, 27, 57, 98 |
|      | C → T | Arg358fs | Takahashi et al. (2009) | WRC45, 46, 48 |
| Hd6  | A → T | 146Lys | Takahashi et al. (2001) | Except for WRC1 |
| Gh7  | T → G | splice_acceptor | | WRC15, 100 |
| RFT1 | G → A | Glu105Lys | Ogiso-Tanaka et al. (2013) | WRC3, 7, 13, 17, 18, 20, 58, 59, 62–65, 98–100 |
| Hd2  | 8-bp deletion | Lys505fs | | WRC7, 9, 10, 16–18, 98, 99 |
|      | T → C | Tyr704His | Koo et al. (2013) | WRC66 |
|      | C → T | Gln705fs | Koo et al. (2013) | WRC2, 31, 38 |
| Hd17 | A → T | Leu132fs | | WRC19 |
|      | A → G | Leu558Ser | Matsubara et al. (2012) | Except for WRC1, 21, 39, 43, 57 |
| DTH8 | 19-bp deletion | Ala32fs | Fujino et al. (2013) | WRC43 |
|      | 1-bp deletion | Lys108fs | | WRC7, 9, 21, 98 |
|      | 8-bp deletion | Glu151fs | | WRC17, 99 |
|      | 1,118-bp deletion | Large deletion | Hori et al. (2015) | WRC3, 10, 19 |

We surveyed mutations in flowering-time genes in the WRC using the TASUKE genome browser. Newly identified mutations found in this study are shown in blue text.

aStop codon.

fs: frameshift.
Fig. 5 GWAS for phenotypes related to agronomical traits and identification of candidate genes for each peak. (A) Manhattan plot for grain color GWAS using a linear mixed model. (B) Manhattan plot for amylose content GWAS using a linear mixed model. (C) Manhattan plot for seed length GWAS using a linear mixed model. Arrows indicate Rc, waxy and GS3 loci. Pink lines represent a P-value threshold corresponding to a false discovery rate of 0.05 using the Benjamini and Hochberg correction.
on chromosome 3 (Fig. 5C, Supplementary Fig. S5) using the MLM and FarmCPU methods. The peak mapped close to GS3 (Os03g0407400), which is known to control the grain length. These mutations in Rc, waxy and GS3 associated with seed traits were previously reported by Huang et al. (2012) in a GWAS with 373 indica accessions and by Wang et al. (2017) in a GWAS with 203 samples from the United States Department of Agriculture Rice Mini-Core Collection (Agrama et al., 2009). The identification of functional variants consistent with previous reports suggested that our compact WRC GWAS population comprising only 69 accessions could match the performance of larger rice core collections.

Finally, we performed GWAS for heading date, which is controlled by a gene network. Significant peaks exceeding a threshold \( P < 0.05 \) with Bonferroni correction for 3,162,968 variants; \( \log_{10}(P_{\text{corr}}) > 7.8 \) were not detected using the MLM or FarmCPU methods (Supplementary Fig. S6). These results implied that the WRC population has limitations for the detection of associations with traits controlled by many QTLs, such as heading date.

Because 11 accessions of the WRC cannot flower in Japanese paddy fields, we excluded these accessions from GWAS for heading date (Supplementary Fig. S5). To investigate the mechanism of the nonflowering phenotype in these accessions, we treated the heading date as a binary trait (Supplementary Table S1). We classified the WRC population as flowering and nonflowering accessions in Japan and coded these data with a numeric value of 1 for flowering accessions and 2 for nonflowering accessions. In GWAS, we detected a significant peak for this numeric data on chromosome 6 using both the MLM and FarmCPU methods and flowering genes were not reported in regions close to this peak (Supplementary Fig. S7). It is possible that new genes located in this region influence flowering time in high latitude areas, such as Japan.

**Discussion**

In this study, we reevaluated WRC accessions using high-resolution WGS data. There is a strong demand to publish the WGS data of the WRC because it is a useful and convenient population for examining natural variation in gene sequences and phenotypes (Sugimoto et al. 2010, Uraguchi and Fujiwara 2013). In addition to the classification of the WRC using WGS data, we performed haplotype analysis of genes in the WRC and developed a GWAS pipeline using the WRC population that could detect candidate genes associated with agronomical traits (Fig. 5). To examine the candidate genes of several traits in detail and perform reverse genetics, we used the TASUKE system to visualize the whole-genome sequence of the WRC for the classification of haplotypes of significant genes among the 69 WRC accessions (Fig. 3, Supplementary Fig. S2).

In our analysis, the WRC was divided into three groups based on their whole-genome sequence variants, broadly consistent with the previous classification using RFLP markers (Kojima et al. 2005). Rice accessions have been previously classified into five groups, indica, aus, temperate japonica, tropical japonica and aromatic using single sequence repeat markers (Garris et al., 2005). However, temperate and tropical japonica were not clearly separated in our analysis using whole-genome sequence (Fig. 1) and it is convenient for analysis to use three main groups for this core collection. Four WRC accessions, WRC20, 22, 24 and 44, originating from the Philippines were classified into the same subgroup (I-4) in the indica group. A small subgroup in aus (A-2) consisted of five accessions originating from Nepal (WRC4, WRC27, WRC29 and WRC30) and Bhutan (WRC28) (Fig. 1A, Supplementary Table S1). In addition, seven of the nine accessions in a small indica subgroup (I-2) were from China or Taiwan (Supplementary Table S1). These results indicated that the classification of WRC using WGS accurately reflects the geographical information of their origin. Large numbers of markers obtained from WGS data may have detected incidental mutations that occur during geographical speciation and made an accurate classification of WRC.

The combination of WGS data and passport data suggests genetic variations in rice cultivar groups. There were several accessions with an early-flowering phenotype in a small indica subgroup (I-2), including several Chinese accessions (Supplementary Table S1). In South China, double cropping of rice is more popular as it increases the rice production (Peng 2014). Seven accessions from South China and Taiwan in the I-2 group have an 8-bp deletion in Hd2 (Table 1) and show relatively early flowering (101–117 d, mean of WRC:122 d, Supplementary Table S1). Hd2 is known as a major determinant of photosensitivity (Gao et al. 2014); therefore, accessions with early flowering and low photosensitivity associated with the 8-bp deletion in Hd2 can be planted at any time of year. These accessions might have been selected with suitable flowering traits for double-cropping rice planting in South China. Among the WRC accessions, there are 11 accessions with a nonflowering phenotype in our experimental field. Eight of these 11 accessions were classified into the largest subgroup in indica (I-1, Fig. 1A). Interestingly, these 11 accessions did not carry any high impact or functional nucleotide polymorphism in the known heading date genes that we examined in this report (Table 1). This result implies that the nonflowering phenotype in WRC accessions in Japan may be regulated by uncharacterized genes.

Because the WRC is a small population for GWAS and contains three groups with highly differentiated accessions, false positives caused by a strong population structure in the panel are a possibility. Therefore, in addition to MLM, we used the statistically robust algorithm, FarmCPU (Liu et al. 2016), to perform GWAS for each trait. We successfully identified Rc, waxy and GS3 genes as genes associated with grain color, amylose content and grain length, respectively (Fig. 5). Of these three traits for which we successfully identified associated gene loci, amylose content and grain color did not exhibit normally distributed phenotypes (Fig. 4). Grain color was measured as a qualitative trait, such as the row type or grain cover in barley, and amylose content was similarly bimodally distributed. Such traits are usually controlled by a small number of genes and are easier to detect in association studies (Milner et al. 2019).
Although grain length in the WRC had an approximately normal distribution (Fig. 4), grain size variation is also often controlled by a few major QTLs (Fan et al. 2006). In conclusion, if there are strong QTLs associated with traits and the number of related QTLs is relatively small, the WRC can be an effective population for detecting causal genes by GWAS. In the case of heading date, we did not detect any significant peaks correlating with well-known heading date genes (Supplementary Fig. S6). The heading date is regulated by a large number of QTLs including several kinds of haplotypes in each locus. Therefore, the small, structured WRC population is likely underpowered for the detection of QTLs associated with complex traits, such as heading date. Using the TASUKE system, we focused on seven heading date genes reported as major QTLs for heading date (Itoh et al. 2018) and identified several high-impact SNPs and indels, such as the introduction of premature stop codons. However, most of these alleles were found in fewer than three accessions in the WRC (Table 1). Because nucleotide polymorphisms with MAF < 0.05 among the 69 accessions were removed from the variant dataset, these rare SNPs were not used in the GWAS procedure. Furthermore, functional mutations in the heading date tended to coincide with population structure. This might be a further reason that known QTLs were not detected in the GWAS for heading date.

Recent reductions in sequencing costs mean that genomic characterization is no longer the main barrier to the exploitation of rice germplasm resources; for most researchers, phenotypic characterization is the most time consuming and costly step. If a GWAS population consists of a large number of accessions and includes only 69 accessions and is a useful set for the deep study of the natural variation in agronomical traits. As we demonstrated in GWAS for seed phenotypes (Fig. 5), the WRC is a compact population and is sufficient to detect the candidate genes reported in previous works using larger numbers of accessions (Huang et al. 2012, Wang et al. 2017). The viability of the WRC for GWAS is meaningful for researchers who have already used and will use the WRC as a phenotyping population. After the detection of associated loci by GWAS, visualization of whole-genome variant data of WRC using TASUKE also helps them to perform an intuitive search for candidate genes.

However, the NARO Genebank maintains > 37,000 rice accessions and not all of the diversity can be represented in a mini-core collection. By selectively generating new short-read sequencing data, expanding the number of accessions with high coverage, whole-genome resequencing data obtained from the genome data from each subgroup of the WRC, we aim to provide a small number of focused populations for GWAS studies based on NARO Genebank materials. The WRC collection can then be used to examine the phenotypic distribution on a trait-by-trait basis, to select an appropriate population for association studies and the identification of unique, agronomically important genes (Fig. 6). Other large collections of rice germplasm are also available, such as the Rice 3K genomes collection (Wang et al. 2018), and we provide here information about the relationship between the WRC and these accessions.
Population structure analysis was conducted using the fastStructure software (Raj et al. 2014) using the Structure_threader wrapper script for parallel implementation (Pina-Martins et al. 2017). The input for fastStructure was the 2,805,329 SNPs and 357,639 indels variant dataset described in the Variant calling. PCA was performed using the R package SNPRelate (Zheng et al. 2012). The input data were 2,805,329 SNPs from the same dataset as above. Principal components were calculated with the snpgdsPCA function.

To compare WRC accessions with the 3K genomes collection, we downloaded a set of 29 million biallelic SNPs with SnpEff annotation from IRRI Snpseek (https://s3.amazonaws.com/3kricegenome/snpseek-dll/NI_bial SNP_pseudo_canonical_ALL.vcf.gz) and intersected with SNPs called for the WRC using bcftools, excluding sites missing from the WRC data, resulting in a dataset of 708,540 SNPs. PCA was performed as above. A phylogenetic tree from 1,725 of these SNPs in approximate linkage equilibrium was constructed using FastTree2 (Price et al. 2010).

Genome-wide association studies

For GWAS, we used MLM models (Yu et al. 2006) and also employed the FarmCPU algorithm (Liu et al. 2016). All association studies were performed using R (R Core Team 2018) using modified scripts from the MVP (https://rrdr.io/github/XiaoleiLiuBio/MVP/), GAPT (Upka et al. 2012) and GENESIS (Gogarten et al. 2019) packages. FarmCPU was used in the parallel implementation by Kusmec and Schnable (2018). Visualization used scripts from MVP, GAPT and the R packages qpmann (Turner 2018) and GWASTools (Gogarten et al. 2012).

Visualization of genotypes using TASUKE

Variants were genotyped by sample up to the GATK HaplotypeCaller (Poplin et al. 2018) step in the Variant calling and filtered using bcftools (Li 2011) with the condition -e ‘QD < 5.0 && FS > 50.0 && QUAL < 5.0 && MQ < 5.0 && MQRankSum < -2.5 && ReadPosRankSum < -1.0 && ReadPosRankSum > 3.5’. Variants for each accession were displayed using the TASUKE genome browser (Kumagai et al. 2013). TASUKE is a web browser-based visualization system for whole-genome variant data. The input files of WRC for TASUKE were created using a custom data analysis pipeline, using the same mapping and variant calls as for the GWAS variant dataset, but without filtering for MAF or allele number. These datasets are accessible at https://ricegenome-corecollection.dna.affrc.go.jp, and users can choose a specific region and/or gene in which they are interested.

Supplementary Data

Supplementary data are available at PCP online.

Acknowledgment

We thank the Advanced Analysis Center of National Agriculture and Food Research Organization (NARO) for the use of facilities.

Funding

NARO Genebank Project to N.T., M.S. and K.E.; Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, QTL5003 to J.W.).

Disclosure

The authors have no conflicts of interest to declare.

References

Agrama, H. A., Yan, W., Lee, F., Fjellstrom, R., Chen, M.-H., Jia, M., et al. (2009) Genetic Assessment of a Mini-Core Subset Developed from the USDA Rice Genebank. Crop Sci. 49: 1336–1346.
Kusmec, A. and Schnable, P.S. (2018) FarmCPUpp: efficient large-scale genomewide association studies. Plant Direct 2: e00053.

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, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Lipka, A.E., Kandianis, C.B., Hudson, M.E., Yu, J., Drnevich, J., Bradbury, P.J., et al. (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr. Opin. Plant Biol. 24: 110–118.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., et al. (2012) CAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Liu, X., Huang, M., Fan, B., Buckler, E.S. and Zhang, Z. (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PloS Genet. 12: e1005767.

Mao, H., Sun, S., Yao, J., Wang, C., Yu, S., Xu, C., et al. (2010) Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. Proc. Natl. Acad. Sci. USA 107: 19579–19584.

Matsubara, K., Ogiso-Tanaka, E., Hori, K., Ebana, K., Ando, T. and Yano, M. (2012) Natural variation in Hd1, a homolog of Arabidopsis ELF3 that is involved in rice photoperiodic flowering. Plant Cell Physiol. 53: 709–716.

Milner, S.G., Jost, M., Taketa, S., Mazón, E.R., Himmelbach, A., Oppermann, M., et al. (2019) Genebank genomics highlights the diversity of a global barley collection. Nat. Genet. 51: 319–326.

Miura, K., Ashikari, M. and Matsukawa, M. (2011) The role of QTLs in the breeding of high-yielding rice. Trends Plant Sci. 16: 319–326.

Norton, G.J., Travis, A.J., Douglas, A., Fairley, S., Alves, E.D.P., Ruang-Areerate, M., et al. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in rice (Oryza sativa), New Phytol. 182: 644–653.

Oraguchi, S. and Fujiwara, T. (2013) Rice breaks ground for cadmium-free cereals. Curr. Opin. Plant Biol. 16: 328–334.

Oraguchi, S., Mori, S., Kuramata, M., Kawasaki, A., Arao, T. and Ishikawa, S. (2009) Root-to-shoot Cd translocation via the xylem is the major process determining shoot and grain cadmium accumulation in rice. J. Exp. Bot. 60: 2677–2688.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinform. 43: 11.10.1–11.10.33.

Wang, W., Maulen, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557: 43–49.

Wang, X., Pang, Y., Zhang, J., Wu, Z., Chen, K., Ali, J., et al. (2017) Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. Sci. Rep. 7:17203.

Wang, Z.-Y., Zheng, F.-Q., Shen, G.-Z., Gao, J.-P., Snustad, D.P., Li, M.-G., et al. (2015) The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. Plant Physiol. 175: 613–622.

Wei, X., Xu, J., Guo, H., Jiang, L., Chen, S., Yu, C., et al. (2010) DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. Plant Physiol. 153: 1747–1758.

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., et al. (2008) Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat. Genet. 40: 761–767.

Yang, M., Lu, K., Zhao, F.-J., Xie, W., Ramakrishna, P., Wang, G., et al. (2018) Genome-wide association studies reveal the genetic basis of ionic variation in rice. Plant Cell 30: 2720–2740.

Yano, M., Katabuye, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., et al. (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell 12: 2473–2483.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P., Hu, L., et al. (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat. Genet. 48: 927–934.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamashita, M., Doebley, J.F., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Zhang, C., Zhu, J., Chen, S., Fan, X., Li, Q., Lu, Y., et al. (2019) Wxlv, the gene controlling rice waxiness, is involved in rice photoperiodic flowering time gene. Proc. Natl. Acad. Sci. USA 107: 5792–5797.

Sweeney, M. T., Thomson, M. J., Cho, Y. G., Park, Y. J., Williamson, S. H., Bustamante, C. D., et al. (2007) Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet. 3: e1133.

Sweeney, M.T., Thomson, M.J., Pfef, 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.

Takahashi, Y., Teshima, K.M., Yokoi, S., Innan, H. and Shimamoto, K. (2009) Variations in Hd1 proteins, Hd3a promoters, and End1 expression levels contribute to diversity of flowering time in cultivated rice. Proc. Natl. Acad. Sci. USA 106: 4555–4560.

Turner, S.D. (2018) qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. J. Open Source Softw 3: 731.

Ueno, D., Kono, I., Yokoshio, K., Ando, T., Yano, M. and Ma, J.F. (2009) A major quantitative trait locus controlling cadmium translocation in rice (Oryza sativa), New Phytol. 182: 644–653.