PCP
PLANT & CELL PHYSIOLOGY

Investigation of the Genetic Diversity of a Rice Core Collection of Japanese Landraces using Whole-Genome Sequencing

Nobuhiro Tanaka[1], Matthew Shenton[1], Yoshihiro Kawahara[1,2], Masahiko Kumagai[2], Hiroaki Sakai[2], Hiroyuki Kanamori[1], Jun-ichi Yonemaru[1], Shinichi Fukuoka[1], Kazuhiko Sugimoto[1], Masao Ishimoto[1], Jianzhong Wu[1] and Kaworu Ebana[3]*

[1]Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
[2]Advanced Analysis Center, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
[3]Genetic Resources Center, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
*Corresponding author: E-mail, ebana@affrc.go.jp; Fax, +81-29-838-7408.

(Received 14 August 2020; Accepted 22 September 2020)

The Rice Core Collection of Japanese Landraces (JRC) consisting of 50 accessions was developed by the genebank at the National Agriculture and Food Research Organization (NARO) in 2008. As a Japanese landrace core collection, the JRC has been used for many research projects, including screening for different phenotypes and allele mining for target genes. To understand the genetic diversity of Japanese Landraces, we performed whole-genome resequencing of these 50 accessions and obtained a total of 2,145,095 single nucleotide polymorphism (SNPs) and 317,832 insertion–deletions (indels) by mapping against the *Oryza sativa* ssp. *japonica* Nipponbare genome. A JRC phylogenetic tree based on 1,394 representative SNPs showed that JRC accessions were divided into two major groups and one small group. We used the multiple genome browser, TASUKE+, to examine the haplotypes of flowering genes and detected new mutations in these genes. Finally, we performed genome-wide association studies (GWAS) for agronomical traits using the JRC and another core collection, the World Rice Core Collection (WRC), comprising 69 accessions also provided by the NARO genebank. In leaf blade width, a strong peak close to *NAL1*, a key gene for the regulation of leaf width, and, in heading date, a peak near *HESO1* involved in flowering regulation were observed in GWAS using the JRC. They were also detected in GWAS using the combined JRC + WRC. Thus, JRC and JRC + WRC are suitable populations for GWAS of particular traits.

Keywords: Genetic diversity • GWAS • Rice core collection • WGS.

Introduction

Exotic materials for breeding can be used to produce advanced rice cultivars, exploiting the variation contained within them to breed for improved performance. To avoid the evaluation of large numbers of samples in large genebank collections of genetic resources, rice core collections have been developed around the world (Kojima et al. 2005, Ebana et al. 2008, Agrama et al. 2009, Zhang et al. 2011). Among the 2,000 Japanese accessions stored in the NARO genebank, 236 accessions were selected based on their passport data. The Rice Core Collection of Japanese Landraces (JRC) comprising 50 accessions was selected to retain 87.5% of the alleles in these 236 accessions based on 32 genome-wide SSR markers (Ebana et al. 2008).

The JRC was developed by the NARO genebank as a suitable population for understanding rice adaptation in northern areas, such as Japan (Ebana et al. 2008).

The JRC and the World Rice Core Collection (WRC), which was also developed by the NARO genebank, have been used for many research projects, in screens for different phenotypes and allele mining for target genes (Suzaki et al. 2009, Ueno et al. 2009, Ochiai et al. 2011, Fujino et al. 2013, Taguchi-Shiobara et al. 2013, Iijima et al. 2019). Because the JRC and WRC include only 50 and 69 accessions, respectively, they are compact and manageable populations for use in these research projects. Although the genetic diversity of JRC seems smaller than that of WRC, significant variations in genes have been identified in the JRC. For example, to identify genes important for the adaptation of rice cultivars to northern latitudes, the JRC accessions were used for the investigation of allelic variations in *Hd5* (Fujino et al. 2013). A 19-bp deletion causing the loss of function of *Hd5* was found in JRC accessions, and it was likely selected as the mutation that causes earlier heading, thus adapting rice to cultivation at higher latitudes. JRC accessions were also used for the evaluation of endosperm enzyme activity to improve the texture and eating quality of rice (Iijima et al. 2019). Evaluation of eating quality and endosperm enzyme activities in a large set of Japanese rice cultivars, including those in the JRC, revealed that low levels of endosperm enzyme activity are associated with high eating quality in rice varieties bred in Japan. BORON EXCESS TOLERANT1 (BET1), responsible for boron-toxicity, was isolated by the observation of cultivar differences among JRC accessions, especially Wataribune (JRC19) (Ochiai et al. 2011). A major cadmium (Cd) transporter, OsHMA3, was

Plant Cell Physiol. 61(12): 2087–2096 (2020) doi:10.1093/pcp/pcaa125, Advance Access publication on 12 October 2020,
available FREE online at https://academic.oup.com/pcp

isolated from WRC30 (Anjana Dhan) through measurements of Cd concentration in shoots of the WRC and JRC (Ueno et al. 2010).

To date, in order to identify causal genes from these core collections, it has been necessary to perform quantitative trait loce (QTL) analyses on F_2 populations derived from crosses between divergent cultivars. High-resolution whole-genome sequencing (WGS) data of core collections enables efficient QTL analysis and gene isolation from cultivars in core collections; however, the JRC WGS data have not been published so far.

Recently, genome-wide association studies (GWAS) have become a common tool in rice, especially because of the reduction in the cost of next-generation sequencing (NGS) analysis. Although GWAS enables us to omit the crossing process for the isolation of causal genes (Yano et al. 2016, Yano et al. 2019), there are a number of factors to consider ineffective GWAS analysis. GWAS populations usually consist of large numbers of landraces and cultivars, e.g. 517 landraces in Huang et al. (2010), 373 *indica* accessions in Huang et al. (2012) and 529 cultivated lines in Yang et al. (2018). Phenotyping of such large populations is time-consuming. Previously, Tanaka et al. (2020) reported that several peaks related to seed traits were detected by using the WRC core collection in GWAS and these peaks were close to well-known genes, such as *Rc*, *waxy* and *GS3*. Although the WRC represents a highly structured population, for some traits, this core collection that consists of only 69 rice cultivars could match the performance of a larger GWAS population (Tanaka et al. 2020).

In this report, we evaluated the genetic diversity of the JRC using WGS data and performed a haplotype analysis of different genes using TASUKE+ to detect new alleles in JRC accessions. We also performed GWAS for normally distributed agronomical traits using the JRC and WRC accessions and detected significant peaks close to well-known genes.

Results

Classification of the JRC based on WGS data

The JRC comprises 50 accessions (**Supplementary Table S1**) and is provided by the genebank at NARO as a suitable population for understanding rice adaptation in northern areas, such as Japan. The NARO WRC was also developed by the genebank at NARO (Kojima et al. 2005) and both core collections have been distributed to >300 research groups for different purposes. To promote effective use of the JRC as a genetic and breeding resource, we obtained WGS data of the 50 accessions and mapped the reads to the *Oryza sativa* ssp. *japonica* Nipponbare genome (Os-Nipponbare-Reference-IRGSP-1.0) (Kawahara et al. 1997) using bwa (Li and Durbin 2009). The average depth was 28.8, ranging from 17.4 to 39.9; the average number of SNPs and insertion–deletions (indels) per accession were 523,212 and 103,299, respectively (**Supplementary Table S1**). When we simultaneously used all 50 JRC accessions, we obtained a total of 2,145,095 SNPs and 317,832 indels by mapping against the *O. sativa* ssp. *japonica* Nipponbare genome. Based on 1394 representative SNPs selected to be in approximate linkage equilibrium (using the Snphylo pipeline; Lee et al. 2014), we constructed a phylogenetic tree to classify the 50 accessions (**Fig. 1A**). JRC accessions were divided into three major groups: two *japonica* groups (J-1 and J-2) and one small *indica* group (I-1). J-2 corresponded to *tropical japonica* varieties, and J-1 was a group of *temperate japonica*. The number of SNPs detected in these accessions is J-1 < J-2 < I-1 reflecting their genetic relationship with the *temperate japonica* reference genome.

Next, we performed joint genotyping of the JRC and WRC accessions. We used a total of 119 accessions and obtained 2,595,145 SNPs and 321,694 indels by mapping against the *O. sativa* ssp. *japonica* Nipponbare genome (Kawahara et al. 1997). Then, we constructed a phylogenetic tree using both the JRC and the WRC accessions based on 2,004 representative SNPs (**Fig. 1B**). In a previous report (Ebana et al. 2008), JRC40, 41, 42, 43 and 44 were classified into the *indica* group based on both 32 SSR markers and their color reaction with phenol; however, JRC21, 41, 42, 43 and 44 were classified into a distant group based on WGS data in this study, while JRC40 was classified as *japonica* (**Fig. 1A, B**).

Haplotype analysis of flowering time genes in the JRC

Heading date is one of the most important agronomical traits, and the diversity of heading date has contributed to the development of rice cultivation in a wider range of latitudes (Khush, 1997). To clarify the functional diversity of the flowering time genes in JRC, we exhibited sequence variation in JRC using a multiple genome browser, TASUKE+ (Kumagai et al. 2013, Kumagai et al. 2019), and performed haplotype analysis for flowering genes (**Fig. 2**, **Table 1**).

We detected three kinds of deletions in *Hd1* among JRC accessions. JRC20, 49, and 50 (Hosogara, Rikutou Rikuu 2 and Himenomochi) carried a 43-bp deletion in exon 1 and JRC14 and 18 carried a 2-bp deletion in exon 2 of *Hd1* (Yano et al. 2000). JRC07 (Iruma Nishiki) carried a 4-bp deletion in exon 1 of *Hd1*, and this deletion has not been previously reported (**Table 1**). A nonsynonymous substitution at amino acid 24 resulting in a premature stop codon was detected in the *Hd1* gene in JRC22 (Mansaku). A nonsynonymous substitution at amino acid 57 resulting in a premature stop codon was also detected in the *Hd1* gene in JRC38 (Nagoya Shiro), and these mutations have also not been previously reported.

JRC44 (Karahoushi) has an 8-bp deletion in the 7th exon of *Hd2*, and JRC41 (Akamai) carries a variant that changes Tyr to His at amino acid 704 in Hd2 CCT domain, resulting in a non-functional protein (Koo et al. 2013) (**Table 1**). JRC21 (Akamai), 42 (Touboshi) and 43 (Akamai) classified into *indica* group have a previously unreported 1-bp deletion in *Hd2* (**Table 1**). These results indicate that haplotype analysis of flowering genes in JRC with TASUKE+ is an efficient way to discover new mutations.

Both *Hd6* and *Hd16*, encoding casein kinase proteins, are involved in photoperiod sensitivity, which is known to be the main determinant of heading date (Takahashi et al. 2001, Hori et al. 2013). In *Hd6*, a nonsynonymous substitution at amino acid 146 resulting in a premature stop codon was observed in JRC23, 27, 30 and 50 (**Table 1**) and the same SNP was also

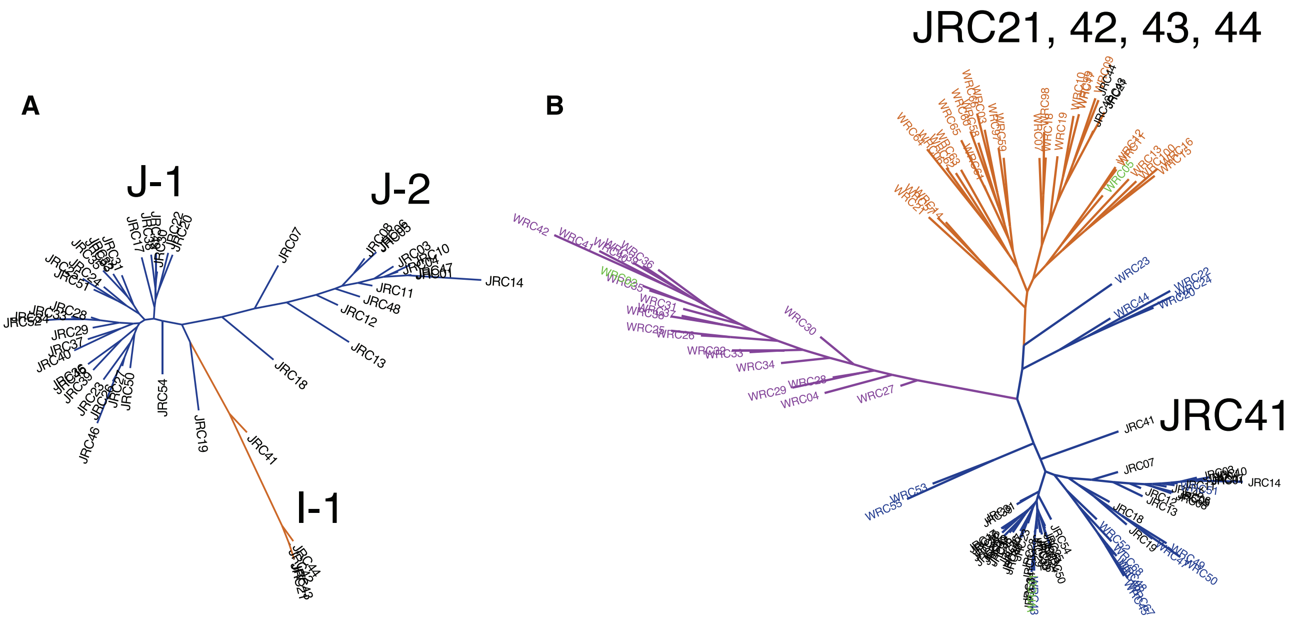
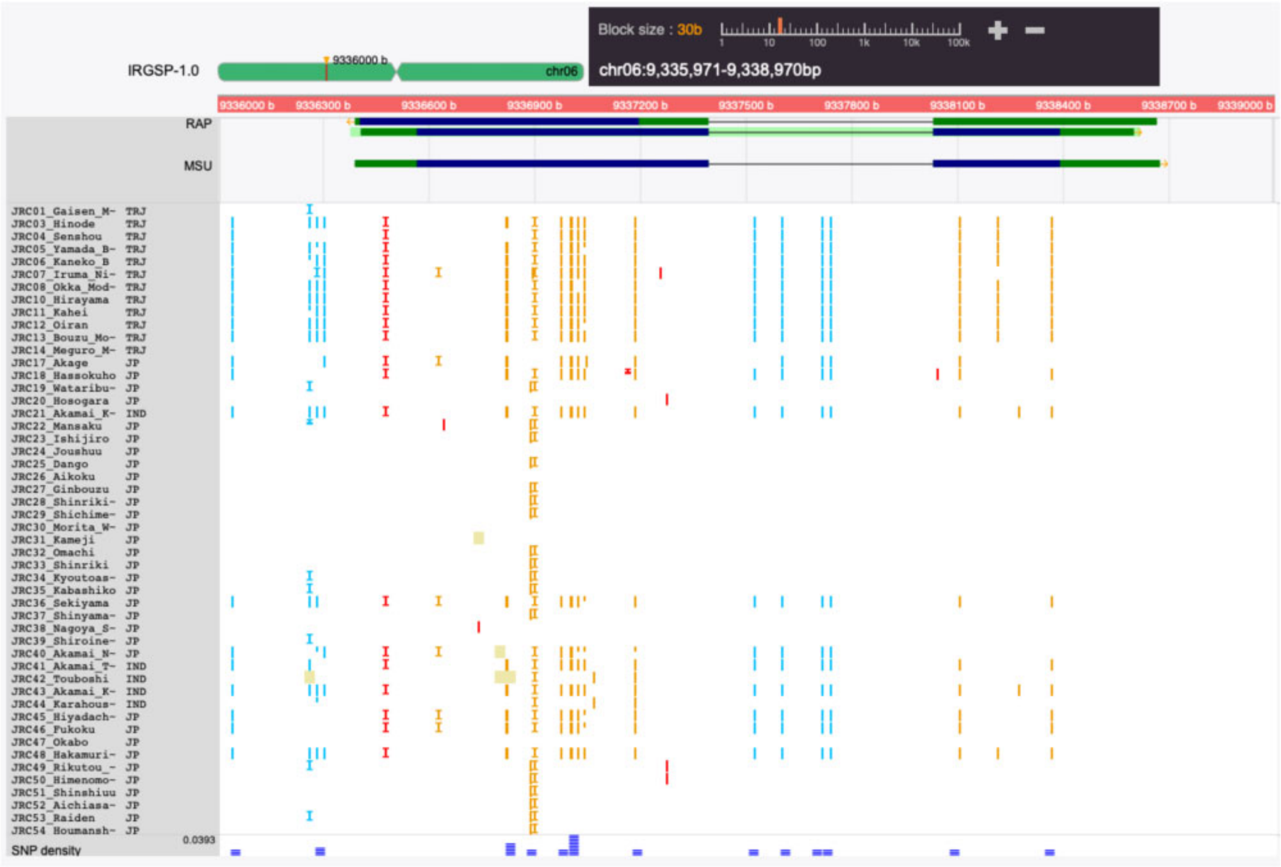


Fig. 1 Phylogenetic tree of JRC (50 accessions) and WRC (69 accessions). (A) Unrooted maximum likelihood tree of the JRC based on 1,394 SNPs called by mapping JRC resequencing data against the *O. sativa* ssp. *japonica* Nipponbare genome. Five accessions, JRC21, 41, 42, 43 and 44, were classified into the *indica* group (red). (B) Unrooted maximum likelihood tree of JRC and WRC based on 2,004 SNPs. JRC and WRC were divided into *japonica* (blue), *indica* (red) and *aus* (magenta). WRC01 (Nipponbare), WRC02 (Kasalath) and WRC05 (Naba) are shown in green color as representative cultivars for each group. JRC accessions are shown in black.



chr06:9,335,971-9,338,970bp

```
JRC01_Gaisen_M-  TRJ
JRC03_Hinode     TRJ
JRC04_Senshou    TRJ
JRC05_Yamada_B-  TRJ
JRC06_Kaneko_B   TRJ
JRC07_Iruma_Ni-  TRJ
JRC08_Okka_Mod-  TRJ
JRC10_Hirayama   TRJ
JRC11_Kahei      TRJ
JRC12_Oiran      TRJ
JRC13_Bouzu_Mo-  TRJ
JRC14_Meguro_M-  TRJ
JRC17_Akage      JP
JRC18_Hassokuho  JP
JRC19_Wataribu-  JP
JRC20_Hosogara   JP
JRC21_Akamai_K-  IND
JRC22_Mansaku    JP
JRC23_Ishijiro   JP
JRC24_Joushuu    JP
JRC25_Dango      JP
JRC26_Aikoku     JP
JRC27_Ginbouzu   JP
JRC28_Shinriki-  JP
JRC29_Shichime-  JP
JRC30_Morita_W-  JP
JRC31_Kameji     JP
JRC32_Omachi     JP
JRC33_Shinriki   JP
JRC34_Kyoutoas-  JP
JRC35_Kabashiko  JP
JRC36_Sekiyama   JP
JRC37_Shinyama-  JP
JRC38_Nagoya_S-  JP
JRC39_Shiroine-  JP
JRC40_Akamai_N-  JP
JRC41_Akamai_T-  IND
JRC42_Touboshi   IND
JRC43_Akamai_K-  IND
JRC44_Karahous-  IND
JRC45_Hiyadach-  JP
JRC46_Fukoku     JP
JRC47_Okabo      JP
JRC48_Hakamuri-  JP
JRC49_Rikutou_-  JP
JRC50_Himenomo-  JP
JRC51_Shinshiuu  JP
JRC52_Aichiasa-  JP
JRC53_Raiden     JP
JRC54_Houmansh-  JP
```

Fig. 2 Screen shot of *Os06g0275000* (*Hd1*) genotypes in JRC accessions displayed using the TASUKE+ system. Each color represents the effect of sequence changes scored by SnpEff. Blue: modifier. Yellow: low. Orange: moderate. Red: high.

Table 1 Distribution of functional mutations in the JRC previously identified in the flowering time genes

QTLs	Indel/SNPs	Substitution	Reference	JRC accessions
Hd1	4 bp Del	Glu230fs		JRC7
	43bp Del	Pro237fs	Yano et al. (2000)	JRC20, 49, 50
	2 bp Del	Phe279fs	Yano et al. (2000)	JRC18
	G → T	Gly24		JRC22
	C → A	Ser57		JRC38
Hd2	1 bp del	Arg82fs		JRC21, 42, 43
	8 bp Del	Lys505fs		JRC44
	T → C	Tyr704His	Koo et al. (2013)	JRC41
Hd16	G → A	Ala331Thr	Hori et al. (2013)	JRC50
Hd6	A → T	146Lys	Takahashi et al. (2001)	Except for JRC23, 27, 30, 50
Ghd7	C → A	Glu53	Xue et al. (2008)	JRC17
RFT1	G → A	Glu105Lys	Ogiso-Tanaka et al. (2013)	JRC21, 42, 43
DTH8	1 bp Del	Lys108fs		JRC21, 43 ,44
Hd17	A → G	Leu558Ser	Matsubara et al. (2012)	JRC1–17, 21, 26–29, 40–44, 46–48, 50, 54

fs, frame shift; , stop codon. Newly identified mutations are shown in blue text.

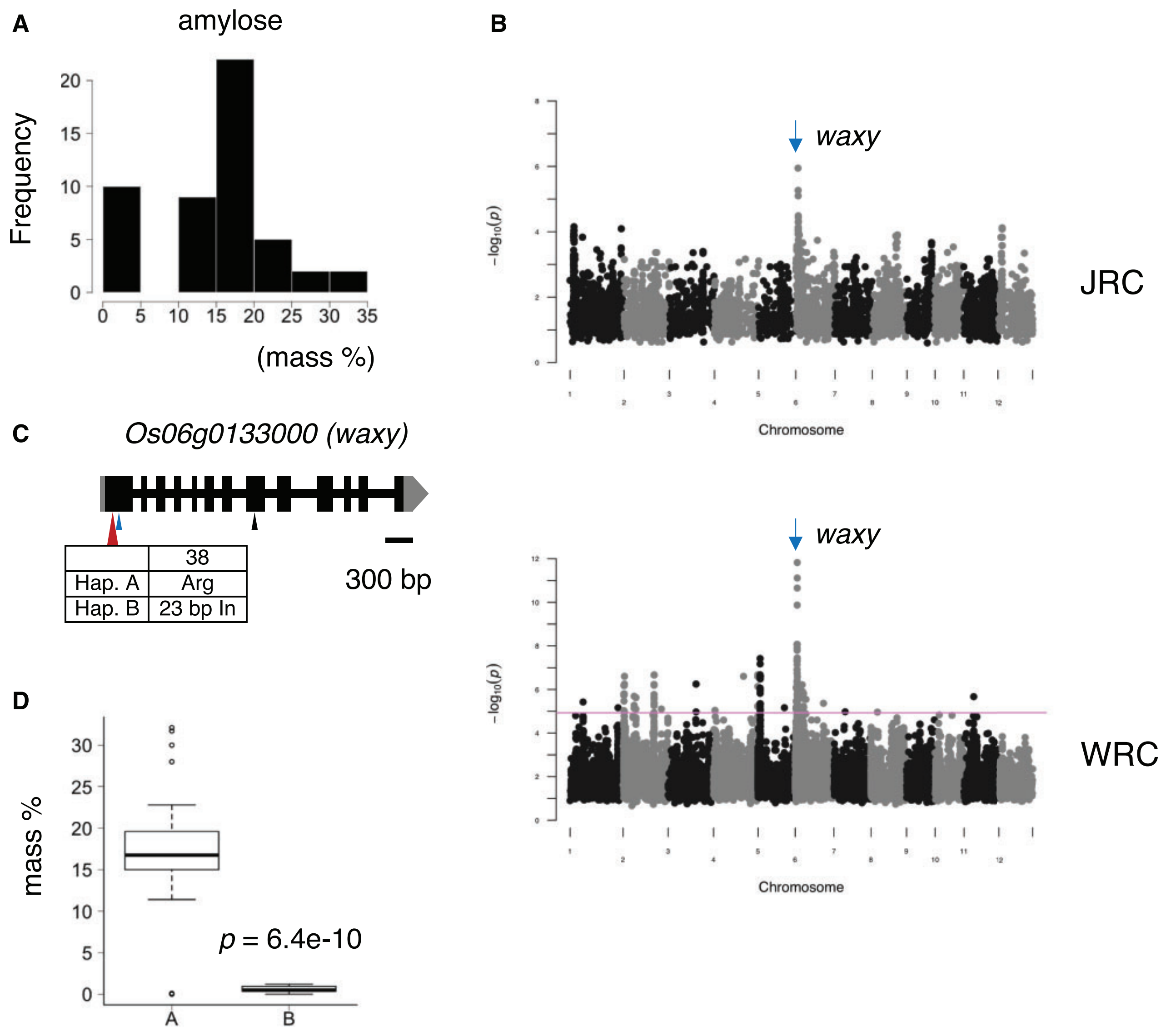
reported in Nipponbare (WRC01) (Takahashi et al. 2001). Only JRC50 carries a functional substitution from Ala to Thr at amino acid 331 of Hd16 which was identified in a commercial Japanese cultivar, Koshihikari (Hori et al. 2013) (**Table 1**). An SNP that changes a Gln to a stop codon at amino acid 53 of Ghd7 (Xue et al. 2008) was only observed in JRC17 (**Table 1**). Matsubara et al. (2012) reported a functional substitution from Leu to Ser at amino acid 558 of Hd17, and this substitution was observed in 27 accessions in the JRC (**Table 1**). JRC21, 42 and 43 have a functional mutation, which changes Glu to Lys at amino acid 105 in RFT1 (Ogiso-Tanaka et al. 2013), and JRC21, 43 and 44 carry a 1-bp deletion in *DTH8* (**Table 1**). JRC21, 42, 43 and 44 were categorized as *indica* in the phylogenetic tree based on WGS data (**Fig. 1**), and these two mutations were also detected only from *indica* accessions in the WRC (Tanaka et al. 2020). We concluded that these mutations occurred after haplogroup differentiation.

GWAS

We performed GWAS for different agronomical traits: amylose content in seed, leaf blade width and heading date, using the JRC and WRC accessions. We compared the utility of the JRC or WRC populations alone, or combined as one JRC + WRC population as a resource for GWAS (**Figs. 3–5**, Supplementary Figs. S1–S3). Previously, GWAS with WRC only detected high impact QTLs, such as *Rc*, *waxy* and *GS3* (Tanaka et al. 2020). First, we used the amylose content of JRC and JRC + WRC seeds for GWAS (**Fig. 3**, **Supplementary Fig. S1**). The distribution of seed amylose content was trimodal (**Fig. 3A**, **Supplementary Fig. S1A**). As with GWAS using the WRC, a significant peak associated with amylose content was observed close to the *waxy* (*Os06g0133000*) gene using the JRC and JRC + WRC populations (**Fig. 3B**, **Supplementary Fig. S1B**). As an outlier (as shown in **Fig. 3D)**, two accessions, JRC22 (Mansaku) and JRC50 (Himenomochi), showed significantly lower amylose content, although they do not have a 23-bp insertion in *waxy* (**Fig. 3C, D**, **Supplementary Table S1**). When we investigated the details of the genome sequence of *waxy* in JRC22 and JRC50, we detected 67-bp deletion in exon 1 of *waxy* in JRC22 and the insertion of a 7.8-kb retrotransposon-like sequence in the *waxy* sequence of JRC50 (Hori et al. 2007). To our knowledge, the 67-bp deletion in the *waxy* gene seen in JRC22 has not been reported before.

Next, we performed GWAS for leaf blade width using JRC only, WRC only and JRC + WRC (**Fig. 4**, **Supplementary Fig. S2**). No significant peak was observed in the result of GWAS with WRC (**Fig. 4B**). JRC and WRC accessions exhibited a normal distribution for leaf width (**Fig. 4A**, **Supplementary Fig. S2A**), and a high peak was observed on chromosome 4 in GWAS using the JRC (**Fig. 4B**). When we used JRC + WRC as the GWAS population, the peak on chromosome 4 was higher than a significance threshold (**Supplementary Fig. S2B**). The detected peaks were close to *NARROW LEAF1* (*NAL1*), *Os04g061500*, known as a major regulator of leaf width (Fujita et al. 2013, Jiang et al. 2015). JRC and WRC accessions carrying haplotype A in *NAL1* had wider leaves than those with haplotype B (**Fig. 4C, D**, **Supplementary Fig. S2C**, D).

Compared with GWAS using the WRC, we improved the detectability of known genes associated with normally distributed phenotypes, such as leaf width, using the JRC only or JRC + WRC as the GWAS population (**Fig. 4B**, **Supplementary Fig. S2B**). Next, we performed GWAS for heading date with JRC only, WRC only and JRC + WRC (**Fig. 5**, **Supplementary Fig. S3**). Because the heading date was affected by a large number of genes (Tsuji et al. 2011), no significant peak was detected by GWAS with WRC only (**Fig. 5B**). On the other hand, relatively high peaks were detected on chromosomes 1, 3 and 11 in the results of GWAS with JRC only (**Fig. 5B**). When we used JRC + WRC as a GWAS population for heading date, these peaks were beyond a *P*-value threshold (**Supplementary Fig. S3B**). On chromosome 1, a peak was located near to the homolog of



	38
Hap. A	Arg
Hap. B	23 bp In

Fig. 3 GWAS for amylose content using the linear mixed model. (A) Histogram of amylose content (mass %) in JRC. (B) Manhattan plot for amylose content using JRC (top) and WRC (bottom) accessions. Arrows indicate the *waxy* locus. (C) Gene structure of and DNA polymorphism in the *waxy* gene. The red arrowhead indicates the position of the 23-bp insertion. The blue arrowhead indicates the position of the 67-bp deletion in JRC22. The black arrowhead indicates the position of the transposon insertion in JRC50. (D) Box plot showing amylose content for JRC accessions bearing haplotype A or haplotype B at the *waxy* gene.

Arabidopsis thaliana HEN1 suppressor 1 (*HESO1*) gene, *Os01g0846450*, recently reported as a new gene for the regulation of days to heading (Yano et al. 2016). JRC and WRC accessions carrying haplotype A in *HESO1* showed earlier heading dates than those with haplotype B (**Fig. 5D**, **Supplementary Fig. S3D**). These results indicate that the JRC and the combined JRC and WRC will be suitable populations for GWAS in particular traits, such as leaf width and heading date.

Discussion

Genetic diversity of the JRC

In this study, we evaluated the genetic diversity of the JRC comparing with the WGS data of the WRC. In addition, we performed haplotype analysis of flowering genes in the JRC using TASUKE+ system and also developed GWAS pipelines using the JRC and WRC accessions that were able to detect known genes for agronomic traits.

Our genetic study revealed that the JRC was divided into three subgroups, J-1, J-2 and I-1 (**Fig. 1**). In passport data recorded when the accessions were collected, JRC01 to JRC14 were upland rice varieties and were also categorized as *tropical japonica*. These 12 accessions in the JRC were included in the J-2 group of our results (**Fig. 1**, **Supplementary Table S1**). These results indicated that the classification of the JRC using WGS accurately reflects their passport data.

Four accessions in the JRC, JRC21, JRC40, JRC41 and JRC43, have the same accession name, Akamai. When we used the WGS data of the JRC, the classification of JRC21 and JRC40

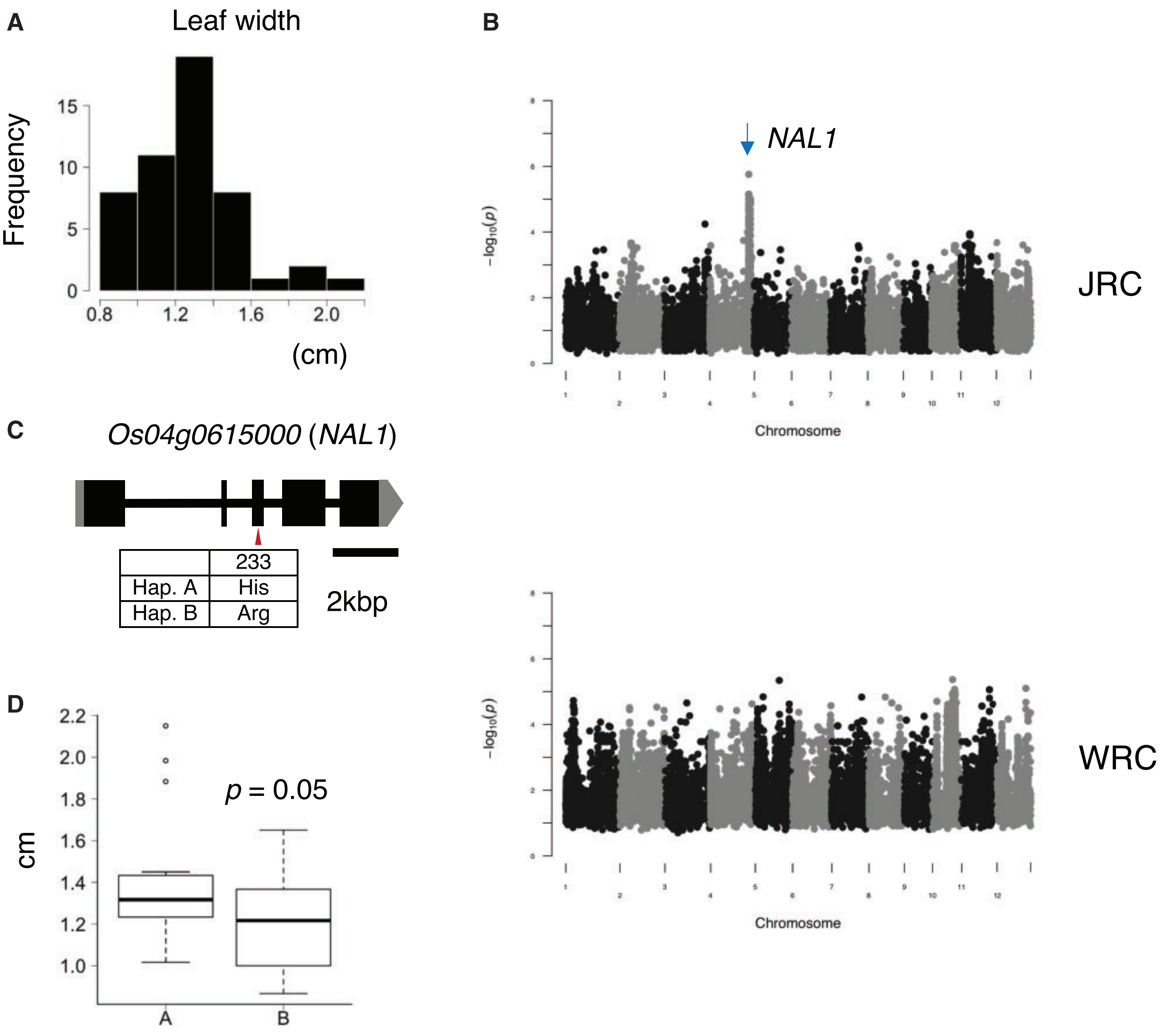


Fig. 4 GWAS for leaf blade width using the linear mixed model. (A) Histogram of leaf blade width in JRC. (B) Manhattan plot for leaf blade width in JRC (top) and WRC (bottom). The arrow indicates the *NAL1* locus. (C) Gene structure of and DNA polymorphism in the *NAL1* gene. The red arrowhead indicates the position of the A → G substitution. (D) Box plot showing leaf blade width for JRC accessions bearing haplotype A or haplotype B at the *NAL1* gene.

was found to be the opposite of that of a previous report where SSR markers were used, i.e. JRC21 was *indica*, not *japonica*, and JRC40 was *japonica*, not *indica*. One advantage of using WGS data to characterize genetic resources is that whole-genome genotype information allows an understanding of their exact genetic diversity, avoiding errors based on small numbers of markers and correcting any errors of identification that may have occurred during collection or propagation.

Haplotype analysis of the *waxy* gene in JRC and WRC accessions

Amylose content in the endosperm is known as an important determinant of rice-eating quality because it reflects the functionality of the rice grain's starch. The *waxy* (*Wx*) gene (*Os06g0133000*) encodes granule-bound starch synthase, which controls amylose synthesis in rice endosperm. The differentiation of glutinous and non-glutinous rice is determined by alleles of the *waxy* locus. Rice cultivars carrying Wx^a have a high (~25–30%) amylose content and those carrying Wx^b, with a G-to-T substitution at the 5′ UTR splicing site, show decreased expression of *waxy*, associated with intermediate (~15–20%) amylose concentration (Hirano et al. 1998) (**Supplementary Table S1**). Glutinous rice varieties generally result from loss-of-function mutations in the *waxy* gene, resulting in very low amylose concentrations. In the JRC and WRC, most glutinous rice cultivars containing little seed amylose have a 23-bp insertion or a retrotransposon-like sequence in *waxy* (Wanchana et al. 2003, Hori et al. 2007). Only JRC22 was glutinous rice without a known allele of *waxy* (**Fig. 3**, **Supplementary Table S1**). A new allele of *waxy* from this JRC accession suggested the usefulness of JRC. This is a small population

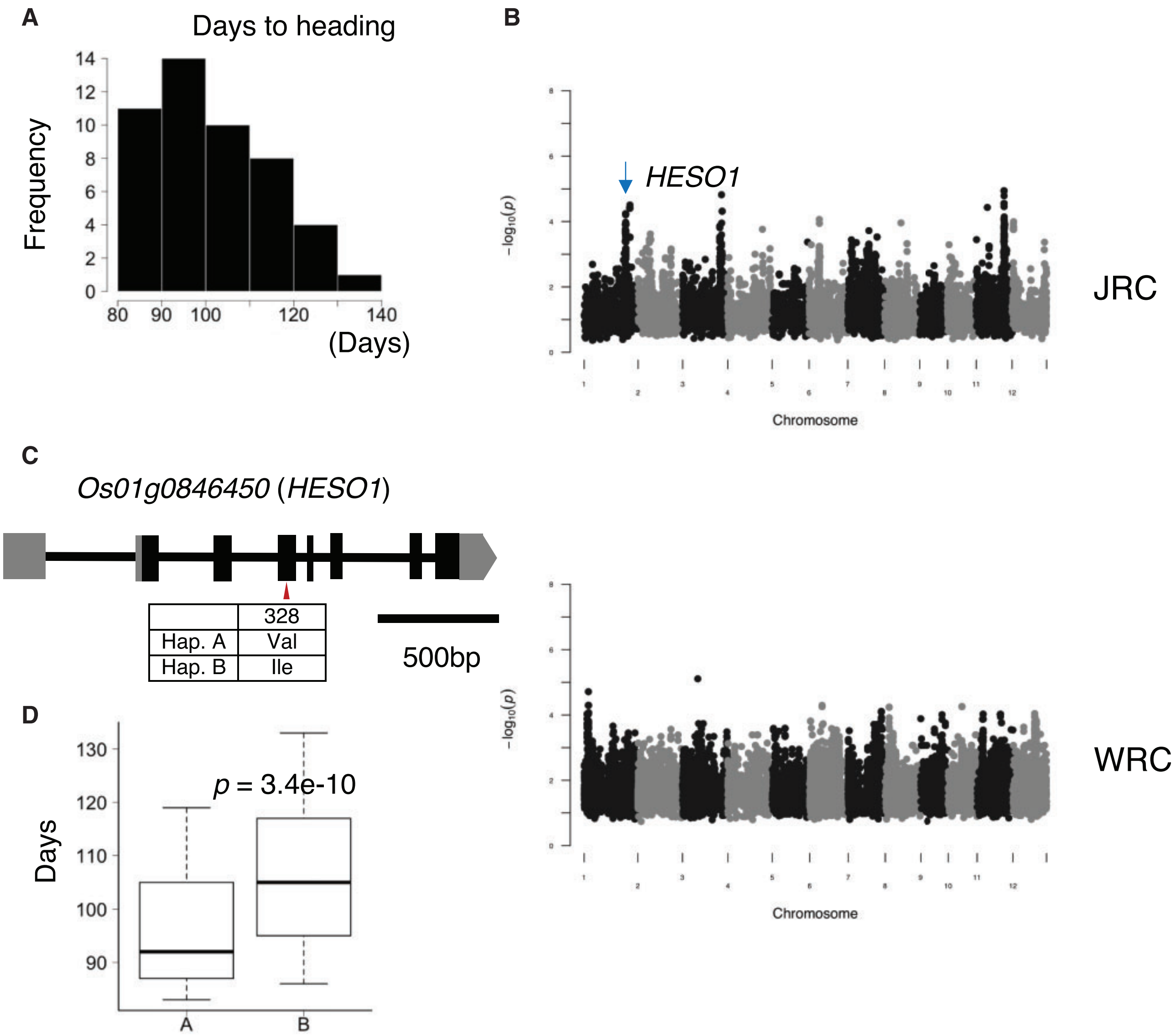


Fig. 5 GWAS for days to heading using the linear mixed model. (A) Histogram of days to heading (days) in JRC. (B) Manhattan plot for days to heading in JRC (top) and WRC (bottom). The arrow indicates the *HESO1* locus. (C) Gene structure of and DNA polymorphism in the *HESO1* gene. The red arrowhead indicates the position of the G → A substitution. (D) Box plot showing days to heading for JRC accessions bearing haplotype A or haplotype B at the *HESO1* gene.

but containing new genetic information useful for rice breeding.

Oryza sativa was derived from ancestral wild rice, *Oryza rufipogon* Griff. *O. rufipogon* carries the *Wx*^{a} allele and the G-to-T substitution at the 5′ UTR splicing site was selected during the domestication of *O. sativa japonica* as the *Wx*^{b} allele (Hirano et al. 1998), resulting in intermediate amylose concentrations, and sticky, but non-glutinous rice. In the JRC and WRC, all glutinous rice accessions carried the *Wx*^{b} allele at the 5′ UTR splicing site, and in addition, the 23-bp insertion, the transposon insertion or the 67-bp deletion inactivating the *waxy* gene and resulting in the glutinous phenotype (**Supplementary Table S1**). These results suggested that glutinous rice was selected from among *japonica* cultivars that were domesticated in East Asia, wherein the *Wx*^{b} allele had become fixed in the population based on the resulting stickiness of the cooked rice. Most Japanese glutinous rice varieties in the JRC obtained the same 23-bp insertion in *waxy*, except for the 67-bp deletion and transposon insertion (**Fig. 3**). These results suggest that the desirability of the glutinous rice phenotype led to the wide distribution of the mutations in the local region.

Genetic diversity of NAL genes in the JRC and WRC

In **Fig. 4**, we detected the *NAL1* locus associated with leaf width using the JRC as a GWAS population, or using the combined JRC + WRC, when the peak was higher than a significance threshold. Haplotype B of *NAL1* was present in almost half of the JRC and WRC accessions; however, no significant peak was observed in GWAS using the WRC alone (**Fig. 4B**). To determine the reason for the difference in GWAS results between the JRC and the WRC, we investigated the relationship between the *NAL1* genotypes and leaf width in the WRC (**Supplementary Fig. S4**). In the WRC, accessions carrying haplotype A in *NAL1* also showed wider leaves than those with haplotype B (**Supplementary Fig. S4A**). These results indicate that population structure and/or kinship relatedness in the WRC might cause a lower *P*-value in *NAL1* locus. Until now, four genes, *NAL1*, *NAL2*, *NAL3* and *NAL7*, have been reported as major regulators of leaf width (Fujino et al. 2008, Ishiwata et al. 2013, Jiang et al. 2015). In the JRC and WRC accessions, we did not find any SNPs in *NAL2* or *NAL3*. In the *NAL7* gene, several moderate-impact nonsynonymous SNPs causing amino acid substitutions were observed in WRC accessions. Conversely, only three accessions categorized into the *indica* group carried a moderate-impact SNP in the JRC (**Supplementary Fig. S4B**). These results suggest that the genetic diversity of *NAL7* gene in the WRC masked the effect of the *NAL1* gene for leaf width. Although the peak was observed in the JRC alone, it became significant using JRC + WRC due to the increased statistical power obtained when using a greater number of accessions.

GWAS for heading date using the JRC and WRC

Although we detected *HESO1* for heading date using the JRC and WRC as the GWAS population, peaks close to strong QTLs for heading dates, such as *Hd1* and *Hd2*, were not observed (**Fig. 5**, **Supplementary Fig. S3**). In *Hd1*, except for the 2-bp deletion in exon 2 (Yano et al. 2000), widely distributed DNA polymorphisms with high impacts, such as indels, were not observed in the JRC using TASUKE+ (**Fig. 2**, **Table 1**). In GWAS with JRC + WRC, DNA polymorphisms with minor allele frequency (MAF) <0.05 among the 119 accessions were removed from the variant dataset. These rare DNA polymorphisms in *Hd1* were not used in the GWAS procedure, so a significant peak close to *Hd1* was not detected in our experiments.

An 8-bp deletion on exon 7 in *Hd2* was commonly observed both in the JRC and WRC; however, it was only distributed in accessions categorized into *indica* group both in JRC and WRC (**Table 1**) (Tanaka et al. 2020). Because of the problem of population structure, the effect of 8-bp deletion in *Hd2* might be ignored in the calculation of GWAS by the linear mixed model (Korte et al. 2012).

Using the JRC and WRC as the GWAS population, a significant peak associated with heading date was detected on chromosome 1 close to *HESO1*. *HESO1* was originally isolated by GWAS with *japonica* cultivars (Yano et al. 2016), and haplotypes A and B were almost equally distributed in the JRC. On the other hand, only WRC01 (Nipponbare, *japonica*) and WRC43 (Dianyu 1, *japonica*) carried the haplotype A allele of the *HESO1* gene among the WRC accessions (**Fig. 5**). In this report, the *japonica* allelic frequency of *HESO1* was sufficient to allow the detection of the *HESO1* locus associated with heading date.

Although JRC contains only 50 accessions, the detectability of GWAS with the JRC for particular traits was higher than with WRC. In addition, the combination of JRC and WRC includes 119 accessions, so it is still possible to collect high-quality phenotype data with low cost and also to improve the statistical power of the detection. For many traits, phenotyping is labor intensive, and thus, it is important to select the minimum size of the population adequate for the detection of significant association peaks. The JRC is a small population without a high degree of population structure except for the five *indica* varieties. In some cases, it could out-perform the WRC in GWAS. However, a better statistical power could be achieved by using the JRC + WRC population if it is possible to acquire phenotype data for all accessions.

Materials and Methods

Plant materials and WGS

We used JRC accessions maintained at the NARO gene bank (**Supplementary Table S1**). Total DNA was extracted from leaves from one plant of each variety using the DNeasy Plant Mini Kit (Qiagen). The DNA libraries were sequenced using Illumina Hiseq 2000 or Hiseq X instruments (Illumina Co, Ltd.), and paired-end reads were obtained. All reads were mapped against Os-Nipponbare-Reference-IRGSP-1.0 (Kawahara et al. 1997) pseudomolecules using bwa mem (Li and Durbin 2009), and duplicates were removed using Picard MarkDuplicates (**http://broadinstitute.github.io/picard/**).

Variant calling

Variants were called essentially following the GATK Best practices for germline SNP/Indel discovery (Auwera et al. 2013). Variants were first called on a by-sample basis using GATK HaplotypeCaller, and then variants were consolidated in a joint calling step with GenotypeGVCFs (Poplin et al. 2017). GATK version 4.0.11.0 was employed for all steps. Variants were then filtered using bcftools view (Li 2011) with the parameters: -m2 -M2 -g hom –output-type z –exclude-uncalled -e "MAF <0.05 || N_MISSING > 0 || QD < 5.0 || FS > 50.0 || SOR > 3.0 || MQ < 50.0|| MQRankSum < -2.5 || ReadPosRankSum < -1.0 || ReadPosRankSum > 3.5", resulting in a set of variants where no position had missing data or a minor allele frequence of <0.05. All nucleotide polymorphisms were categorized for their potential effects using SnpEff 4.3t (Cingolani et al. 2012) with the Oryza_sativa database.

Phylogenetic tree and population structure

The Snphylo pipeline (Lee et al. 2014) was used to create a maximum likelihood phylogenetic tree based on representative genomic SNPs. The pipeline was employed using default parameters and 100 bootstrap replicates to create the bootstrapped maximum likelihood tree.

GWAS

For GWAS, we used MLM models (Yu et al. 2006). Association studies were performed using R (R Core Team 2018) using modified scripts from the MVP (**https://rdrr.io/github/XiaoleiLiuBio/MVP/**), GAPIT (Lipka et al. 2012) and GENESIS (Gogarten et al. 2019) packages. Visualization used scripts from MVP, GAPIT and the R package qqman (Turner 2014). We removed variants with MAF < 5% from the relevant dataset in GWAS when using the WRC, JRC or JRC + WRC populations.

Visualization of genotypes using TASUKE+

Variants were genotyped by sample up to the GATK HaplotypeCaller (Poplin et al. 2017) step in the variant calling section as above and filtered using bcftools (Li 2011) with the condition -e "QD < 5.0 || FS > 50.0 || SOR > 3.0 || MQ < 50.0 || MQRankSum < -2.5 || ReadPosRankSum < -1.0 || ReadPosRankSum > 3.5". Variants for each accession were displayed using the TASUKE+ genome browser (Kumagai et al. 2019). 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 minor allele frequency 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.

Accession number

The sequence data have been deposited in the DNA Data Bank of Japan Sequence Read Archive (DRA). The accession numbers are from **DRR240814** to **DRR240859** and **DRR095338**, **DRR095344**, **DRR095348**, and **DRR095346** (**Supplementary Table S1**).

Supplementary Data

Supplementary data are available at PCP online.

Funding

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

Acknowledgment

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

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.

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

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6: 80–92.

Ebana, 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.

Fujino, K., Matsuda, Y., Ozawa, K., Nishimura, T., Koshiba, T., Fraaije, M.W., et al. (2008) NARROW LEAF 7 controls leaf shape mediated by auxin in rice. *Mol. Genet. Genomics* 279: 499–507.

Fujino, K., Yamanouchi, U. and Yano, M. (2013) Roles of the Hd5 gene controlling heading date for adaptation to the northern limits of rice cultivation. *Theor. Appl. Genet.* 126: 611–618.

Fujita, D., Trijatmiko, K.R., Tagle, A.G., Sapasap, M.V., Koide, Y., Sasaki, K., et al. (2013) NAL1 allele from a rice landrace greatly increases yield in modern indica cultivars. *Proc. Natl. Acad. Sci. USA* 110: 20431–20436.

Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., et al. (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 35: 5346–5348.

Hirano, H.Y., Eiguchi, M. and Sano, Y. (1998) A single base change altered the regulation of the Waxy gene at the posttranscriptional level during the domestication of rice. *Mol. Biol. Evol.* 15: 978–987.

Hori, K., Ogiso-Tanaka, E., Matsubara, K., Yamanouchi, U., Ebana, K. and Yano, M. (2013) Hd16, a gene for casein kinase I, is involved in the control of rice flowering time by modulating the day-length response. *Plant J.* 76: n/a–46.

Hori, Y., Fujimoto, R., Sato, Y. and Nishio, T. (2007) A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (*Oryza sativa*). *Theor. Appl. Genet.* 115: 217–224.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., 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., et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44: 32–39.

Iijima, K., Suzuki, K., Hori, K., Ebana, K., Kimura, K., Tsujii, Y., et al. (2019) Endosperm enzyme activity is responsible for texture and eating quality of cooked rice grains in Japanese cultivars. *Biosci. Biotechnol. Biochem.* 83: 502–510.

Ishiwata, A., Ozawa, M., Nagasaki, H., Kato, M., Noda, Y., Yamaguchi, T., et al. (2013) Two WUSCHEL-related homeobox genes, narrow leaf2 and narrow leaf3, control leaf width in rice. *Plant Cell Physiol*. 54: 779–792.

Jiang, D., Fang, J., Lou, L., Zhao, J., Yuan, S., Yin, L., et al. (2015) Characterization of a null allelic mutant of the rice NAL1 gene reveals its role in regulating cell division. *PLoS One* 10: e0118169.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R.

Khush, G. S. (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol. 35: 25–34.

Ouyang, S., et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y). 6.

Kojima, Y., Ebana, K., Fukuoka, S., Nagamine, T. and Kawase, M. (2005) Development of an RFLP-based rice diversity research set of germplasm. *Breed. Sci.* 55: 431–440.

Koo, B.-H., Yoo, S.-C., Park, J.-W., Kwon, C.-T., Lee, B.-D., An, G., et al. (2013) Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Mol. Plant* 6: 1877–1888.

Korte, A., Vilhjálmsson, 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.

Kumagai, M., Kim, J., Itoh, R. and Itoh, T. (2013) Tasuke: a web-based visualization program for large-scale resequencing data. *Bioinformatics* 29: 1806–1808.

Kumagai, M., Nishikawa, D., Kawahara, Y., Wakimoto, H., Itoh, R., Tabei, N., et al. (2019) TASUKE+: a web-based platform for exploring GWAS results and large-scale resequencing data. *DNA Res.* 26: 445–452.

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

Matsubara, K., Ogiso-Tanaka, E., Hori, K., Ebana, K., Ando, T. and Yano, M. (2012) Natural variation in Hd17, a homolog of Arabidopsis ELF3 that is involved in rice photoperiodic flowering. *Plant Cell Physiol.* 53: 709–716.

Ochiai, K., Shimizu, A., Okumoto, Y., Fujiwara, T. and Matoh, T. (2011) Suppression of a NAC-Like transcription factor gene improves boron-toxicity tolerance in rice. *Plant Physiol.* 156: 1457–1463.

Ogiso-Tanaka, E., Matsubara, K., Yamamoto, S., Nonoue, Y., Wu, J., Fujisawa, H., et al. (2013) Natural variation of the RICE FLOWERING LOCUS T 1 contributes to flowering time divergence in rice. *PLoS One* 8: e75959.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., et al. (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL **https://www.R-project.org/**.

Suzaki, T., Ohneda, M., Toriba, T., Yoshida, A. and Hirano, H.-Y. (2009) FON2 SPARE1 redundantly regulates floral meristem maintenance with FLORAL ORGAN NUMBER2 in rice. *PLoS Genet.* 5: e1000693.

Taguchi-Shiobara, F., Ozaki, H., Sato, H., Maeda, H., Kojima, Y., Ebitani, T., et al. (2013) Mapping and validation of QTLs for rice sheath blight resistance. *Breed. Sci.* 63: 301–308.

Takahashi, Y., Shomura, A., Sasaki, T. and Yano, M. (2001) Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the α subunit of protein kinase CK2. *Proc. Natl. Acad. Sci. USA* 98: 7922–7927.

Tanaka, N., Shenton, M., Kawahara, Y., Kumagai, M., Sakai, H., Kanamori, H., et al. (2020) Whole-genome sequencing of the NARO world rice core collection (WRC) as the basis for diversity and association studies. *Plant Cell Physiol*. 61: 922–932.

Tsuji, H., Taoka, K. and Shimamoto, K. (2011) Regulation of flowering in rice: two florigen genes, a complex gene network, and natural variation. *Curr. Opin. Plant Biol*. 14: 45–52.

Turner, S.D. (2014) qqman: an R package for visualizing GWAS results using Q–Q and manhattan plots, doi: 10.1101/005165.

Ueno, D., Kono, I., Yokosho, 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.

Ueno, D., Yamaji, N., Kono, I., Huang, C.F., Ando, T., Yano, M., et al. (2010) Gene limiting cadmium accumulation in rice. *Proc. Natl. Acad. Sci. USA* 107: 16500–16505.

Wanchana, S., Toojinda, T., Tragoonrung, S. and Vanavichit, A. (2003) Duplicated coding sequence in the waxy allele of tropical glutinous rice (Oryza sativa L.). *Plant Sci*. 165: 1193–1199.

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 ionomic variation in rice. *Plant Cell* 30: 2720–2740.

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.

Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., et al. (2019) GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl. Acad. Sci. USA* 116: 21262–21267.

Yano, M., Katayose, 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.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, 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, P., Li, J., Li, X., Liu, X., Zhao, X. and Lu, Y. (2011) Population structure and genetic diversity in a rice core collection (*Oryza sativa* L.) investigated with SSR markers. *PLoS One* 6: e27565.