Genome-wide association study of rice grain width variation
Xiao-Ming Zheng, Tingting Gong, Hong-Ling Ou, Dayuan Xue, Weihua Qiao, Junrui Wang, Sha Liu, Qingwen Yang, and Kenneth M. Olsen

Abstract: Seed size is variable within many plant species, and understanding the underlying genetic factors can provide insights into mechanisms of local environmental adaptation. Here we make use of the abundant genomic and germplasm resources available for rice (Oryza sativa) to perform a large-scale genome-wide association study (GWAS) of grain width. Grain width varies widely within the crop and is also known to show climate-associated variation across populations of its wild progenitor. Using a filtered dataset of >1.9 million genome-wide SNPs in a sample of 570 cultivated and wild rice accessions, we performed GWAS with two complementary models, GLM and MLM. The models yielded 10 and 33 significant associations, respectively, and jointly yielded seven candidate locus regions, two of which have been previously identified. Analyses of nucleotide diversity and haplotype distributions at these loci revealed signatures of selection and patterns consistent with adaptive introgression of grain width alleles across rice variety groups. The results provide a 50% increase in the total number of rice grain width loci mapped to date and support a polygenic model whereby grain width is shaped by gene-by-environment interactions. These loci can potentially serve as candidates for studies of adaptive seed size variation in wild grass species.

Key words: grain size, genome-wide association study (GWAS), General Linear Model (GLM), Mixed Linear Model (MLM), Oryza sativa, rice.

Introduction

Seed size has long been recognized as an important contributor to habitat-specific adaptation in plants (Chapin et al. 1993). Production of larger seeds can provide fitness advantages in unfavorable environments, where the probability of seedling establishment is low, but it often comes at the cost of reduced total seed production (Westoby et al. 1992; Leishman 2001). Studies of seed size variation in diverse species have revealed a complex developmental and genetic basis, with phenotypes in natural populations determined to varying degrees by environmental variation, maternal effects, heritable genetic variation, and developmental constraints (Westoby et al. 1992; Gnan et al. 2014). Understanding the extent to which this important life history trait is under genetic control, and identifying the genes that control it, can provide insights into the genetic basis of adaptation across spatially heterogeneous or temporally varying environments.

Received: 23 May 2017. Accepted: 23 October 2017.

Corresponding Editor: Loretta Johnson.

X.-M. Zheng.* Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100081, P.R. China; Department of Biology, Campus Box 1137, Washington University, St. Louis, MO 63130, USA.
T. Gong.* Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100081, P.R. China; Department of Life and Environmental Science, Minzu University of China, Beijing, 100081, P.R. China.
H.-L. Ou. Department of Clinical Laboratory, The General Hospital of PLA Rocket Force, Beijing, 100875, P.R. China.
D. Xue. Department of Life and Environmental Science, Minzu University of China, Beijing, 100081, P.R. China.
W. Qiao, J. Wang, S. Liu, and Q. Yang. Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100081, P.R. China.
K.M. Olsen. Department of Biology, Campus Box 1137, Washington University, St. Louis, MO 63130, USA.

Corresponding authors: Kenneth M. Olsen (email: kolsen@wustl.edu); Qingwen Yang (email: yangqingwen@caas.cn).

*These authors contributed equally to this work.
1This paper is part of a Special Issue entitled Ecological Genomics.
Genomic model species can provide a powerful resource for detailed genetic characterization of seed size variation. Among grass species, domesticated rice (Oryza sativa L.) features a well-annotated reference genome and a large foundation of knowledge on the genetic and developmental basis of seed size variation, including molecularly characterized candidate genes identified through forward and reverse genetics (Fan et al. 2006; Song et al. 2007; Shomura et al. 2008; Weng et al. 2008; Takano-Kai et al. 2009; Li et al. 2011). While the range of phenotypic variation in domesticated rice primarily reflects selection at the hands of humans rather than natural selection, the underlying genetic bases for grain size variation can nonetheless provide key insights into the genetic architecture of this ecologically important trait. Genes controlling seed size in rice can also serve as candidates for exploring the genetic basis of adaptive seed size variation in wild grass species.

In the context of crop domestication, grain size was one of the key agronomic traits targeted for selection during the history of rice cultivation. Larger grain size is an important factor for increased crop yield (Fuller 2007; Lu et al. 2013). Interestingly, the increases in this trait that occurred in the earliest stages of rice domestication may not have been for increased yield per se, but rather a side effect of unintentional selection for seeds that could survive cultivation-associated burial under the soil surface (Purugganan and Fuller 2009). Grain morphology is a quantitative trait and can be described as three elements: grain length, width, and thickness (Xing and Zhang 2010). Among these elements, grain width was likely subject to deliberate selection earlier than the others (Fuller 2007).

Besides affecting crop yield and seed survival in the soil, rice grain size and shape are also important determinants of cooking and taste qualities (Sun et al. 2013), and because of this, selection during domestication was not universally in the direction of increased size. A range of grain morphologies can thus be found in present-day rice varieties that far exceeds that of the wild ancestor, O. rufipogon Griff. In general, longer and more slender rice grains tend to be preferred in South and Southeast Asia, southern China, the USA, and Latin America, while shorter, rounder grains are preferred in northern China, Korea, Japan, and parts of the Mediterranean (Vaughan et al. 2008; Fuller 2012). Grain morphology is also correlated with the genetic subgroups that are present within the two subspecies of rice, O. sativa. subsp. indica and O. sativa subsp. japonica. There are two major genetic subgroups within the indica subspecies (aus and indica), and these tend to have longer and more slender grains than the three subgroups within the japonica subspecies (tropical japonica, temperate japonica, and aromatic varieties).

As an important agronomic trait in rice and other cereal crops, grain morphology and its genetic basis have been examined in numerous studies, and more than 400 quantitative trait loci (QTLs) have been described in rice in recent decades (Thomson et al. 2003; Aluko et al. 2004; Huang et al. 2010, 2012; Zhao et al. 2011). In only a handful of these cases have the underlying genes been cloned and molecularly characterized; these include GW2 (Song et al. 2007), qSW5 (Shomura et al. 2008), GS5 (Li et al. 2011), GS8 (Wang et al. 2012), and GS6 (Sun et al. 2013). GW2 was among the first rice grain size genes to be cloned and was reported to have a large effect on grain width (Song et al. 2007); however, the 1-bp causal deletion in this gene has subsequently been found to be relatively rare in cultivated rice varieties (Lu et al. 2013). Investigations of genetic diversity and molecular evolution of the gene qSW5 showed that it was strongly selected in japonica cultivars, potentially for increased crop yield (Sun et al. 2013; Shomura et al. 2008). The genes GS5 and GS6 underlie minor QTLs for grain width whose function is masked by qSW5 (Lu et al. 2013). For GW8, accesses of Basmati rice (a widely cultivated South Asian variety within the aromatic subgroup of subsp. japonica) were found to carry loss-of-function haplotypes that reduce grain filling and confer the long and slender grains that are characteristic of this variety; in contrast, high-yielding indica cultivars examined in the same study did not carry this loss-of-function variation (Wang et al. 2012).

Besides genetic control, rice grain size and shape can also be strongly influenced by environmental factors. For example, an increase in growing temperature from 21 to 30 °C was found to result in a 3.5% reduction in grain size in a study of indica varieties (Cao et al. 2009), and elevated nighttime temperatures can have a particularly major effect in reducing grain width (Cheng et al. 2009). Interestingly, this effect of temperature on grain size development has been documented not only in domesticated rice but also in populations of its wild ancestor. In a study of phenotypic variation in natural and transplanted populations of O. rufipogon at different locations across China, Zhou et al. (2013) reported larger grain size development at higher latitudes, an effect that was largely attributable to the effects of temperature on plant growth.

To date, characterizations of the genetic basis of grain size and shape variation in rice have mostly relied on linkage mapping populations derived from biparental crosses. While useful for identifying the allelic variation that differs between two phenotypically distinct parents, this approach does not sample the larger pool of genetic variation that may contribute to phenotypic variation within a species. In contrast, genome-wide association studies (GWAS) using populations of unrelated individuals can potentially reveal a larger and more representative set of loci that contribute to phenotypic variation. As demonstrated by a number of recent studies (Huang et al. 2010; Zhao et al. 2011; Yano et al. 2016), GWAS in rice can allow fast and efficient identification of important loci for domestication-related traits and their molecular bases. The effectiveness of this approach depends on dense genetic marker coverage of the sort that can be obtained through whole-genome sequencing, assembly, and alignment; it is thus most amenable to genomic model species such as rice with well-characterized reference genomes. For species like rice where phenotypic variation is correlated with genetic subgroups, GWAS approaches must also be able to effectively detect significant phenotypic associations through the background noise created by population structure (stratification).

In this study, we employed GWAS in a large sample of cultivated rice varieties (397 accesses) to examine the genetic basis of grain morphological variation, with a specific focus on grain width, using analyses designed to be robust to population structure. Our results reveal several newly identified grain width candidate loci that can serve as the focus of future studies to characterize the molecular bases of the observed variation, and they demonstrate the power of GWAS when combined with whole genome sequence data and large, genetically diverse sample sets.

Materials and methods
Selection of plant materials and grain phenotyping
A total of 570 accesses were used in the study, including 256 O. sativa subsp. indica accesses (189 indica and 67 aus varieties), 141 O. sativa subsp. japonica accesses (85 temperate japonica, 46 tropical japonica, and 10 aromatic varieties), and 173 wild rice accesses (O. rufipogon) collected from China, South Asia, and Southeast Asia. Accession details are provided in the supplementary data, Table S1. Cultivated rice accesses were selected from core germplasm resources of China and other countries to repre-
analyzed in FaST-LMM. This method retains the computational
method is specifically designed to handle association mapping for
able software package, GAPIT (Tang et al. 2016). The SUPER
implemented using the SUPER method in R in the publicly avail-
resulting from
consisting of 100 000 randomly sampled SNPs from the full SNP
structure was performed using STRUCTURE (v2.3.4) with a dataset
samples excluded; and (4) SNPs with minor allele frequency <5%
least 10 bp; (3) SNPs with missing data in >20% of cultivated rice
et al. (2010, 2012): (1) bi-allelic SNPs only; (2) SNPs separated by at
extracted DNA from fresh leaves of greenhouse-grown plants us-
sequencing was performed on the Illumina HiSeq-
ncgr.ac.cn/12chr/index.asp. For the remaining 227 accessions, we
unmapped and non-unique reads (Li et al. 2009).
SNP detection was performed using SAMtools, and SNPs were
called at the population level using the cultivated rice samples
regions were compared to cultivated rice in the analyses of se-
quences were called at the population level using the cultivated rice samples
width was measured at the widest point for each caryopsis
harvested from each cultivated rice accession upon maturity, and
grain width (ANOVA; 
P< 0.0001), with most of the varia-
minor allele frequency <5%.

GWAS and analysis of mapped loci
Association analysis to identify loci controlling grain width was
implemented using the SUPER method in R in the publicly avail-
able software package, GAPIT (Tang et al. 2016). The SUPER
method is specifically designed to handle association mapping for
large genetic marker datasets and is therefore appropriate for the
>1.9 million SNP dataset analyzed here. This method proceeds by
extracting subsets of SNPs from the total dataset, which are
analyzed in FaST-LMM. This method retains the computational
advantage of FaST-LMM, and also increases statistical power even
when compared to using the entire set of SNPs by other methods
(C.H. Wang et al. 2014). It is well suited for association mapping in
sample sets with population structure and thus is appropriate for
GWAS in rice (Zhang et al. 2016). To improve the accuracy, we
adopted two models, GLM (General Linear Model) and MLM (Mixed
Linear Model); these were implemented using MLM-SUPER and
GLM-SUPER. Manhattan and quantile-quantile (Q-Q) plots were
generated using the R package qqman (Turner 2014). To reduce
the false-positive rate but also retain major associations, we tested
different P-value thresholds combining a false discovery rate (FDR)
at P < 0.05 and Bonferroni correction to choose thresholds based
on Q-Q plots and locations of known grain width loci.

Genotyping and population structure analysis
Using whole genome sequences from 397 cultivated rice acces-
sions, we generated a filtered dataset consisting of 1 922 258 SNPs
having a minor allele frequency (MAF) greater than 5%. SNP calls
at these loci were then used to assess population structure, genetic
diversity, and associations with grain width variation. STRUCTURE analysis indicated support for K = 2 populations as assessed by 
\( \Delta K \) (Evanno et al. 2005; Fig. S1B); however, strong
support for K = 2 may be an artifact of rejecting extremely low
likelihoods for K = 1 (Vigouroux et al. 2008), so we also considered
higher population values up to K = 6. This revealed a second opti-
mum at K = 5, where each genetic subpopulation was assigned
primarily to a single variety group (indica, aus, aromatic, tropical
japonica, temperate japonica) in a pattern consistent with previous
studies of rice population structure (Caicedo et al. 2007; Huang
et al. 2012) (Figs. S1A, S1B). Population assignments for individ-
ual accessions at K = 5 closely matched results of a previous
STRUCTURE analysis that included these accessions (Huang et al.
2012; Wang et al. 2013); differences were mostly in the assign-
ment of a few accessions to either the tropical japonica or closely related
temperate japonica subgroup (see Table S1 supplementary table for membership coeffi-
cient values).

Grain width variation and genome-wide associations
Grain width varied widely among rice varieties, ranging from a
minimum of 1.10 mm in one aus accession to 2.80 mm in a temperate
japonica accession (mean = 1.97 mm, standard deviation = 0.18, N =
397; Table S1). The five rice subgroups differed significantly in
grain width (ANOVA; F = 15.2, P < 0.0001), with most of the varia-
tion distributed between members of the indica and japonica
subspecies (t-test, P < 0.0001; Fig. S2).

To maximize statistical power in the GWAS, we used the SUPER
(Settlement of MLM Under Progressively Exhaustive Relationship)
method, which divides the whole genome into smaller bins and
selects the subset of influential bins to identify association signals in the full population (M. Wang et al. 2014). Manhattan plots of grain width associations are shown in Fig. 1 for the two models employed (MLM and GLM; Figs. 1A and 1C); a higher-resolution plot for chromosome 5 is shown in Fig. S32. Q-Q plots indicate that both models fit the data fairly well (Figs. 1B and 1D; see also Huang et al. 2012).

Following Bonferroni correction, the cutoff for statistical significance was determined to be $P < 2.6 \times 10^{-8}$. For the MLM model, this threshold yielded 33 significant peak-like signals across the genome; these include six regions that have been previously identified as associated with grain width variation (Table S22). For the GLM model, 10 significant peak-like signals were detected, including two loci identified in previous studies; seven of the loci identified by GLM were also identified by the MLM model (Table S22).

To provide a discrete subset of loci for further examination, we focused subsequent analyses on the set of seven loci identified by both methods. These occur on chromosomes 2, 4, and 5 (Table 1). SNP associations with grain width at these seven loci are shown in Fig. 2. Two of the candidate loci, both located in intergenic regions of chromosome 5 (C5-4.7, C5-5.4), occur within 200 kb of loci that were previously identified to be associated with grain width variation (Huang et al. 2010; Zhao et al. 2011); one of these, C5-5.4, is in close proximity to the genomic region containing the well-characterized grain width locus $qSW5$ (Shomura et al. 2008; Huang et al. 2010; Zhao et al. 2011). For all seven candidate loci, the SNP variants predominating in subsp. japonica accessions (i.e., tropical japonica, temperate japonica, and aromatic varieties) were associated with increased grain width.

Table 1. Genome-wide significant signals for grain width identified by the combined MLM and GLM models using the SUPER method.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Candidate locus position (IRGSP 1.0)</th>
<th>Major allele</th>
<th>Minor allele</th>
<th>Minor allele freq.</th>
<th>$P$ (MLM)</th>
<th>$P$ (GLM)</th>
<th>SNP location</th>
<th>SNP locations in previous studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>C2-11.7 11705678 C G</td>
<td>0.37</td>
<td>7.76E–10</td>
<td>1.51E–08</td>
<td>Os02G0301800 intron —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>C4-11.8 11816702 C T</td>
<td>0.29</td>
<td>2.75E–12</td>
<td>1.05E–08</td>
<td>Os04G0278900 upstream —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C5-1.7 1660867 T A</td>
<td>0.34</td>
<td>1.44E–08</td>
<td>3.06E–10</td>
<td>Os05G0128200 upstream —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C5-2.7 2704165 T A</td>
<td>0.23</td>
<td>2.38E–10</td>
<td>7.67E–08</td>
<td>Os05G0147100 exon —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C5-4.7 4694469 G A</td>
<td>0.39</td>
<td>2.97E–08</td>
<td>3.77E–09</td>
<td>Intergenic, 4.7 kb upstream of Os05G0178600 —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C5-5.4 5371916 G T</td>
<td>0.34</td>
<td>5.58E–08</td>
<td>9.83E–08</td>
<td>Intergenic, 6.8 kb upstream of Os05G0187500 —</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C5-19.4 19375934 C T</td>
<td>0.33</td>
<td>2.87E–09</td>
<td>1.48E–09</td>
<td>Os05G0398450 exon —</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Genomic locations of previously detected grain width QTLs that occur within 200 kb of the loci identified in the present study.

*Gene names are indicated for SNPs located within 2 kb of an annotated gene.

*Genes reported in previous studies.
SNP distributions in rice subgroups and evidence for adaptive introgression

Table 2 presents information on the distributions of the candidate locus SNPs across the five rice variety subgroups. For three of the seven loci (C2-11.7, C5-1.7, C5-19.4), SNP distributions are congruent with indica–japonica phylogenetic divergence, with the increased grain width variant predominantly in japonica rice and largely absent in subsp. indica accessions. These patterns are also evident in haplotype trees for each candidate locus region (Figs. S4A–S4F). At two of the loci (C5-2.7, C5-4.7), the increased grain width variant occurs at high frequency in only a subset of subsp. japonica varieties; this pattern could potentially reflect selection for wider grains in specific japonica varietal subgroups (e.g., temperate japonica at C5-2.7; see Fig. S4D). At the remaining two loci, C4-11.8 and C5-5.4 (qSW5), large subsets of subsp. indica accessions carry the japonica SNP (Figs. S4B, S4F). This distribution is potentially consistent with selective introgression of japonica alleles into subsp. indica varieties. Interestingly, there is little overlap between these two loci in the particular subsp. indica accessions that carry the japonica variant. At C4-11.8, the japonica SNP is universally present in aus accessions and occurs in 40% of indica varieties; in contrast, at C5-5.4 (qSW5), the japonica SNP is present in <5% of aus varieties but more than two-thirds of indica accessions. The japonica variant of C5-5.4 (qSW5) is especially well represented in Chinese indica varieties, where it occurs at a frequency of >95%; by comparison, less than 3% of these Chinese accessions carry the C4-11.8 japonica variant.

Nucleotide diversity and deviations from neutral equilibrium at candidate loci

Estimates of nucleotide diversity and associated Tajima’s D values for the seven candidate locus regions are presented in Table 3 (see also Table S3). Average nucleotide diversity across the loci was lower in domesticated rice than in its wild progenitor, consistent with domestication bottlenecks and (or) selection; there was a 28.7% reduction in the average π value across loci and a 43.9% reduction in the average value of θw. Consistent with the previously documented population structure present within domesticated rice, all seven loci show positive Tajima’s D values for O. sativa, with statistically significant deviations at P < 0.0001 occurring at six of the seven loci. Similarly, positive Tajima’s D values occur in O. rufipogon at all loci, with significant deviations at P < 0.0001 for three of the loci.

More interesting deviations from neutral equilibrium are present within specific rice subgroups that may reflect selection on grain width. Two of the loci, C2-11.7 and C5-4.7, show significantly negative Tajima’s D values for subsp. indica overall or subgroups therein. These deviations are consistent with an excess of low-frequency polymorphisms within subsp. indica caused by the presence of a few accessions that carry introgressed japonica haplotypes (see Figs. S4A–S4F). Thus, the negative deviations from neutrality at these two loci may reflect selection for wider grain width in a few indica accessions which has occurred through adaptive introgression of japonica alleles. For C4-11.8 and C5-5.4 (qSW5), the much larger proportion of indica accessions that carry japonica haplotypes at these loci (see above) generates significantly positive Tajima’s D values due to the resulting haplogroup structure with associated deep branches (Figs. S4B, S4F). The locus C5-5.4 (qSW5) is further characterized by significantly negative deviations within subsp. japonica; this could potentially reflect an episode of positive selection for a favored wider-grain japonica allele at this previously identified grain width locus

Sun et al. 2013; Shomura et al. 2008.)
### Table 3.

<table>
<thead>
<tr>
<th>Species/Group</th>
<th>C5-4.7</th>
<th>C5-5.4</th>
<th>C5-19.4</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oryza sativa</em></td>
<td>3.6166</td>
<td>3.3959</td>
<td>3.9301</td>
</tr>
<tr>
<td><em>Oryza indica</em></td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td><em>Oryza japonica</em></td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**Note:** *N* = average number of nucleotide differences per site between two sequences (Niu 1997) calculated on the total number of polymorphic sites; *P* < 0.05; **P** < 0.01; ***P*** < 0.001.

---

### Discussion

Identifying the genes that control seed size variation can provide important insights into the genetic basis of adaptation, including local adaptation across environmentally heterogeneous species ranges. Here we have made use of the abundant grain size variation that is present in domesticated rice, together with the germplasm and genomic resources that are available for this genomic model crop species (Huang et al. 2010, 2012), to perform GWAS and map loci associated with grain width variation. GWAS based on large-scale resequencing data provides a powerful platform to map genetic variants that underlie phenotypic diversity (Morris et al. 2013; Huang and Han 2014). Using the statistically powerful SURF method, which is designed to be robust to population substructure such as is found in rice (Q, Wang et al. 2014), we identified 33 loci by the MLM and 10 loci by the GLM model, with seven loci detected by both models. Previous GWAS studies in rice have identified a total of 10 significant grain width loci, which include two of those identified in the present analysis (C5-4.7 and C5-5.4; Table 2; Huang et al. 2010; Zhao et al. 2011). Thus, five of the loci jointly detected through the MLM and GLM models are newly identified, and together these represent a 50% increase in the total number of grain width loci that have now been described. Between the two models employed in this study, GLM is less stringent than MLM (Price et al. 2015) whereas MLM can over-fit a model and create type II errors (Xue et al. 2013). Thus, using both methods applied in conjunction can be considered preferable to using them individually.

It is interesting to note that only one of the seven loci identified through the joint GLM/MLM analysis, C5-5.4 (qSW5), corresponds to one of the grain width genes previously identified through fine-mapping in a QTL mapping population (Shomura et al. 2008). QTL mapping has traditionally relied on populations derived from biparental crosses. As such, the phenotypic and genetic variation within the mapping population represents only a tiny fraction of the total variation present within a species. While such limited sampling may not matter for traits controlled by just one or two major-effect loci, it inevitably under-samples the genetic contributors for polygenic traits like grain width. GWAS, by comparison, can yield a more representative set of loci controlling phenotypic variation within a species, provided the population sampling is representative of the species overall. However, limited statistical power in detecting loci can be a more acute problem for GWAS, particularly if there is population structure and/or the rate of linkage disequilibrium decay is high relative to marker density. Thus, it is likely that the GWAS conducted here and in previous rice studies (Huang et al. 2010; Zhao et al. 2011) have failed to detect some of the loci affecting grain width variation. The 15 grain width loci identified to date through GWAS (Table 1; Huang et al. 2010; Zhao et al. 2011) are best considered as a candidate subset of the total pool of genetic contributors.

A particular challenge in mapping grain width in rice is that this phenotype is closely correlated with the genetic subgroups present within the species; varieties within the japonica subspecies tend to have wider grains than those within the indica subspecies (Table S1; Fig. S2). Association mapping methods must therefore possess the statistical power to detect associated loci despite the confounding effects of this population structure (Yu et al. 2006; Price et al. 2010). None of the candidate loci mapped here shows a perfect correlation between SNP distributions and the indica–japonica subspecies divergence (Table 2), as any such SNP would have been filtered out when controlling for population structure. Our ability to successfully identify grain width loci (Table S2), including 50% more loci than identified in previous studies (Huang et al. 2010; Zhao et al. 2011), supports the effectiveness of the SUPER method in handling large SNP datasets such as the >1.9 million SNPs examined here. The significant deviations from neutrality that we detect for haplotype sequences at several
of these loci (Table 3) further suggests that variation within or near these candidate loci is causally related to grain width variation.

Among the seven candidate loci, two of them, C4-11.8 and C5-5.4 (qSWS), show nucleotide and haplotype distributions that are most clearly suggestive of selective introgression of japonica alleles into subsp. indica varieties (Tables 2, 3; Figs. S4B, S4F). Similarly, Takano-Kai et al. (2009) documented that a mutation in the gene G5S conferring increased grain length originated in subsp. japonica, with subsequent introgression into subsp. indica. These alleles would have been transferred between the rice subspecies through the process of introgressive hybridization, which was likely facilitated by the higher outcrossing rates among early rice cultivars and by the physical proximity of the two subspecies as a result of human migration throughout Asia (Kovach et al. 2009; Vaughan et al. 2008; Izawa et al. 2009; Huang et al. 2012).

The fact that different sets of indica accessions carry the putatively introgressed japonica alleles at the two loci is consistent with two independent selective events. Interestingly, the phenotypic effect of the japonica allele also appears to differ at the two loci. At C4-11.8, the japonica allele shows a trend that is consistent with it conferring wider grains; while not statistically significantly different, mean grain width for those indica accessions with the japonica allele is 1.927 compared to 1.911 for those with the indica allele. However, the trend is in the opposite direction at C5-5.4, where mean grain width for indica accessions with and without the japonica allele is 1.910 and 1.923, respectively. This lack of a clear correlation between candidate locus SNPs and their predicted phenotypes suggests that the genetic architecture of grain width is more complex than one of simple additivity across contributing loci. This observation is consistent with previous studies of grain width loci, where the effects of the minor-effect G5S and G5E QTLs were shown to be masked by qSWS (Lu et al. 2013). Follow-up studies would be helpful for further characterizing the nature of gene regions affecting rice grain width.

Grain width measurements in the present study were made using greenhouse-grown plants, which allowed us to control for potential environmental effects. It is interesting to note, however, that for rice plants cultivated in the field, the climates in which indica and japonica rice varieties are farmed would be expected to augment the patterns of grain width variation observed here. In general, varieties in the indica subspecies tend to be grown in hotter climates at lower latitudes; japonica varieties are grown in more moderate climates, with temperate japonicas grown in cooler conditions than tropical japonicas (Fuller 2012). Since warmer growing environments promote the development of more slender grains, both in cultivated rice and its wild progenitor (Cooper et al. 2008; Cao et al. 2009; Cheng et al. 2009; Zhou et al. 2013), this geographical trend would be expected to broaden the phenotypic range beyond what we observed in the greenhouse. Moreover, since grain shape is also closely correlated with cooking qualities and taste, local cultural preferences have likely further shaped the distributions of wider-grain and slender-grain rice varieties across Asia (Fuller et al. 2009). Regional variation in grain width may thus reflect a combination of trade-offs between environmental adaptation, crop yield, and cultural preferences.

While the GWAS performed here focused exclusively on cultivated varieties of a domesticated crop species, the findings of this study also have implications for seed morphology and adaptive variation in nondomesticated species. One area in which these data have very direct applicability is in understanding the mechanisms of invasiveness and local adaptation in weedy rice populations, which have evolved multiple times during the history of rice cultivation and which aggressively outcompete crop varieties in rice production areas worldwide (e.g., Xia et al. 2011; Li et al. 2017). For grain width, the combination of a polygenic basis and climatically contingent developmental plasticity may be especially likely to promote the spread of invasive strains, since there are apparently few constraints on this trait that might otherwise prevent locally adapted phenotypes from arising (see, e.g., Zenni et al. 2014). The grain width loci mapped here and in previous studies could also have direct applicability for exploring the genetic components of local adaptation across the range of O. rufipogon (Zhou et al. 2013) and other wild species of Oryza. For more distantly related grasses, these loci can, at a minimum, provide a point of comparison for examining the genetic architecture of adaptive seed size variation; or they could, with further molecular characterization, potentially serve as candidate genes for understanding the molecular basis of adaptive seed size variation.

Conclusions

This study presents an in-depth survey of the genetic association and diversity of grain width in a large panel of genetically and geographically diverse rice germplasm from across Asia. The phenotyping methods and mapping resolution presented in this study points to the existence of numerous QTLs that merit further dissection, potentially involving expression-QTL mapping or molecular studies using targeted genome editing. By identifying, understanding, and integrating subpopulation-specific variation using a combination of approaches, the loci identified here may prove useful to rice breeders and efforts to close the gap between grain development and yield optimization in rice, as well as to those interested in adaptive seed variation in wild grass species. Future work will focus on using recombinant inbred lines (RILs) and (or) nested association mapping (NAM) together with functional genomics techniques to validate the effects of these genes and their functional variants.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2016YD1000301) and the Innovation Project of the Chinese Academy of Agricultural Sciences (Collection and Introduction of Crop Germplasm Resources). X.-M.Z. is grateful to the China Scholarship Council for funds supporting her university visit as a research scholar in the Olsen Lab at Washington University. We thank members of the Olsen Lab group for helpful comments on the manuscript.

References


