NRC Research Press

ARTICLE

Genome-wide association study of rice grain width variation[1]

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.

Résumé : La taille des graines est variable chez de nombreuses espèces végétales et comprendre les facteurs génétiques qui sous-tendent cette variation peut fournir un éclairage sur les mécanismes d'adaptation aux conditions locales. Dans ce travail, les auteurs emploient les abondantes ressources génomiques et génétiques disponibles chez le riz (*Oryza sativa*) pour réaliser une analyse d'association pan-génomique (GWAS) à grande échelle sur la largeur des graines. La largeur des graines varie grandement au sein de cette espèce et présente une variation associée au climat parmi les populations du riz sauvage. À l'aide d'un jeu filtré de >1,9 million de marqueurs SNP à travers le génome chez 570 accessions de riz cultivé et sauvage, les auteurs ont réalisé un GWAS à l'aide de deux modèles complémentaires, GLM et MLM. Les modèles ont livré 10 et 33 associations significatives, respectivement, et ont conjointement révélé sept régions candidates, dont deux avaient été identifiées précédemment. Des analyses de la diversité nucléotidique et de la distribution des haplotypes à ces locus ont révélé des signatures de sélection et une structure suggérant une introgression adaptative d'allèles pour la largeur des graines parmi les différents groupes du riz. Ces résultats augmentent de 50% le nombre total de locus pour la largeur des graines chez le riz et supportent un modèle polygénique où la largeur des grains est déterminée par des interactions gène-environnement. Ces locus pourront possiblement servir de candidats pour de futures études sur la variation adaptative au sein des espèces sauvages de graminées. [Traduit par la Rédaction]

Mots-clés : taille des graines, analyse d'association pan-génomique (GWAS), modèle linéaire général (GLM), modèle linéaire mixte (MLM), *Oryza sativa*, riz.

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 (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.

[1]This 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), *GW8* (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*, accessions 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 accessions) 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 accessions were used in the study, including 256 *O. sativa* subsp. *indica* accessions (189 *indica* and 67 *aus* varieties), 141 *O. sativa* subsp. *japonica* accessions (85 *temperate japonica*, 46 *tropical japonica*, and 10 *aromatic* varieties), and 173 wild rice accessions (*O. rufipogon*) collected from China, South Asia, and Southeast Asia. Accession details are provided in the supplementary data, Table S1[2]. Cultivated rice accessions were selected from core germplasm resources of China and other countries to repre-

[2]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2017-0106.

sent all five major genetic subgroups within the crop based on previous analyses by Huang et al. (2012) and Wang et al. (2013). These include accessions of two rice variety groups not examined in previous association studies of grain width variation (*aus* and *aromatic* rice). A sample of 173 *O. rufipogon* accessions was selected to represent the geographical distribution of the wild progenitor. The wild accessions were not phenotyped for grain width, as seed production is more sporadic in wild rice, but their haplotype sequences were compared to cultivated rice in the analyses of sequence variation focusing on candidate loci identified by GWAS.

Seeds of all domesticated rice accessions were planted and grown in the greenhouse at the Chinese Academy of Agricultural Sciences (CAAS) for up to two years to provide material for genotyping and grain phenotyping. A random sample of grains was harvested from each cultivated rice accession upon maturity, and grain width was measured at the widest point for each caryopsis (unhulled grain) using an electronic digital caliper. Three independent measurements were made per accession, with the mean value used in association mapping.

DNA sequencing and generation of the filtered SNP dataset

Whole genome sequences have been previously published for 343 of the selected accessions, and sequence data for these were downloaded from the following databases: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/ERX/ and http://www.ncgr.ac.cn/12chr/index.asp. For the remaining 227 accessions, we extracted DNA from fresh leaves of greenhouse-grown plants using a modified CTAB method (Murray and Thompson 1980). Whole genome DNA sequencing was performed on the Illumina HiSeq-2000 platform by Novogene Corporation (Beijing, China). Paired-end resequencing reads were mapped to the Nipponbare rice reference genome using BWA (V0.6.1-r104) with the default parameters (Li and Durbin 2009). SAMtools (V0.1.18) software was used to convert mapping results into BAM format and to filter the 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 and the SAMtools mpileup package (Li et al. 2009). Differences from the rice reference genome were called as candidate SNPs. The resulting calls were then filtered using the following four criteria to remove substandard arrays and SNPs, following Huang et al. (2010, 2012): (1) bi-allelic SNPs only; (2) SNPs separated by at least 10 bp; (3) SNPs with missing data in >20% of cultivated rice samples excluded; and (4) SNPs with minor allele frequency <5% excluded. This filtering approach yielded a total of 1 922 258 SNPs used for the genetic analyses described below.

To confirm that sampled rice accessions represent the five major subgroups within the crop, a Bayesian analysis of population structure was performed using STRUCTURE (v2.3.4) with a dataset consisting of 100 000 randomly sampled SNPs from the full SNP dataset. We calculated the genetic component for each accession using a burn-in length of 10 000 followed by a run length of 50 000 based on the admixture model. The ΔK ad hoc statistic of Evanno et al. (2005) was used to assess assumed population numbers ranging from $K = 2$ to 6.

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 available 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.

For each grain width candidate region identified in the GWAS, we used the SNPs within the region to construct Neighbor-Joining haplotype trees in MEGA version 6 (Tamura et al. 2013); tree construction was performed with the Jones–Taylor–Thornton (JTT) model with gaps/missing data excluded. For associated SNPs that mapped to intergenic regions or putative gene promoter regions, we included polymorphisms in a 2.0-kb region surrounding the marker SNP with highest statistical association with the phenotype. For associated SNPs within open reading frames (ORFs), we included variation in the region encompassing the entire coding sequence. Wild rice sequences were included in haplotype tree construction for comparison to the cultivated rice haplotypes. We performed bootstrap analysis (1000 replicates) to assess branch support; however, the low number of polymorphic sites within the sequenced regions resulted in low statistical support for most branches, even in the absence of homoplasy, and bootstrap results are not presented. Measures of nucleotide diversity (π, Nei (1987); θ_w, Watterson (1975)) and deviations from neutral equilibrium (Tajima's *D*, Tajima 1989) were calculated for each candidate gene region, using DnaSP v5.10.01 (Rozas 2009) to identify potential signatures of selection at the grain width candidate loci.

Results

Genotyping and population structure analysis

Using whole genome sequences from 397 cultivated rice accessions, 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 ΔK (Evanno et al. 2005; Fig. S1B[2]); 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 optimum 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[2], S1B[2]). Population assignments for individual 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 assignment of a few accessions to either the *tropical japonica* or closely related *temperate japonica* subgroup (see Table S1[2] for membership coefficient 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[2]). The five rice subgroups differed significantly in grain width (ANOVA; $F = 15.2$, $P < 0.0001$), with most of the variation distributed between members of the *indica* and *japonica* subspecies (*t*-test, $P < 0.0001$; Fig. S2[2]).

To maximize statistical power in the GWAS, we used the SUPER (Settlement of MLM Under Progressively Exclusive Relationship) method, which divides the whole genome into smaller bins and

Fig. 1. Genome-wide association studies (GWAS) of grain width. (A) Manhattan plots of the MLM-SUPER model. (B) Quantile-quantile (Q-Q) plot of the MLM-SUPER model. (C) Manhattan plots of the GLM-SUPER model. (D) Q-Q plot of the GLM-SUPER mode. Negative log-transformed *P* values from a genome-wide scan are plotted against position on each of 12 chromosomes. Red horizontal dashed line indicates the genome-wide significance threshold after Bonferroni correction. [Colour online.]

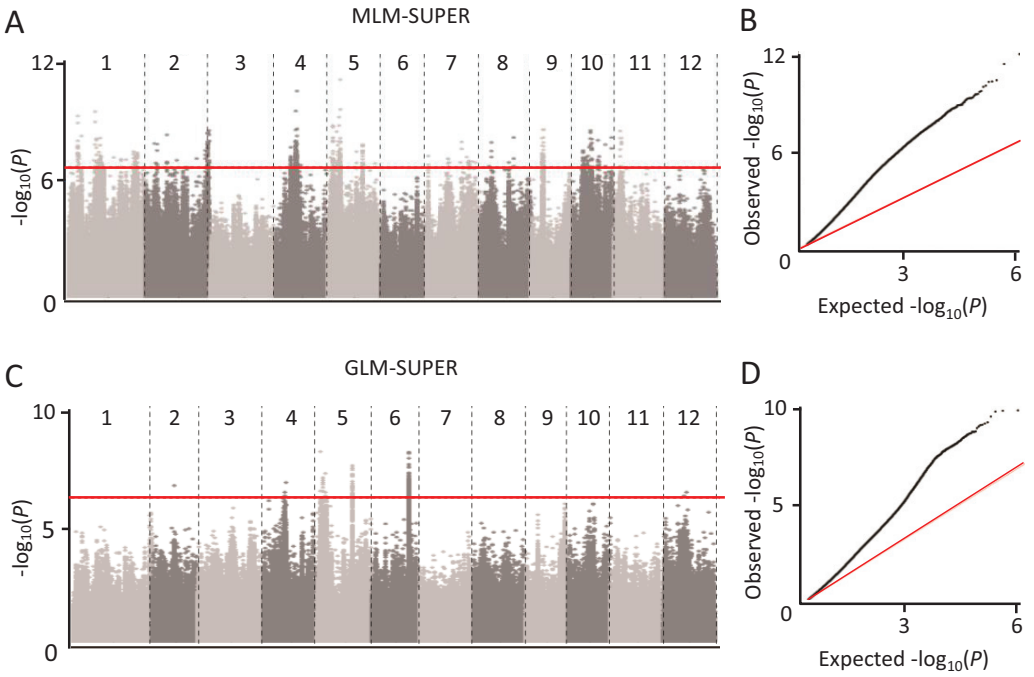

Table 1. Genome-wide significant signals for grain width identified by the combined MLM and GLM models using the SUPER method.

Chr.	Candidate locus	Genome position (IRGSP 1.0)	Major allele	Minor allele	Minor allele freq.	*P* (MLM)	*P* (GLM)	SNP location	SNP locations in previous studies[a]
2	C2-11.7	11705678	C	G	0.37	7.76E–10	1.51E–08	Os02G0301800 intron	—
4	C4-11.8	11816702	C	T	0.29	2.75E–12	1.05E–08	Os04G0278900 upstream[b]	—
5	C5-1.7	1660867	T	A	0.34	1.44E–08	3.06E–10	Os05G0128200 upstream[b]	—
5	C5-2.7	2704165	T	A	0.23	2.38E–10	7.67E–08	Os05G0147100 exon	—
5	C5-4.7	4694469	G	A	0.39	2.97E–08	3.77E–09	Intergenic, 4.7 kb upstream of Os05G0178600[c]	4850026 (Huang et al. 2010)
5	C5-5.4	5371916	G	T	0.34	5.58E–08	9.83E–08	Intergenic, 6.8 kb upstream of Os05G0187500[c]	5359253 (Huang et al. 2010) 5343859 (Zhao et al. 2011)
5	C5-19.4	19375934	C	T	0.33	2.87E–09	1.48E–09	Os05G0398450 exon	—

[a]Genomic locations of previously detected grain width QTLs that occur within 200 kb of the loci identified in the present study.
[b]Gene names are indicated for SNPs located within 2 kb of an annotated gene.
[c]Genes reported in previous studies.

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. S3[2]. 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 S2[2]). 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 S2[2]).

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.

Fig. 2. SNP associations with grain width at seven candidate loci identified through genome-wide association studies (GWAS). Bars indicate mean values; black lines indicate standard errors. ***, mean values are significantly different at $P < 0.0001$ for all loci. Higher frequency nucleotides are on the left side for each locus.

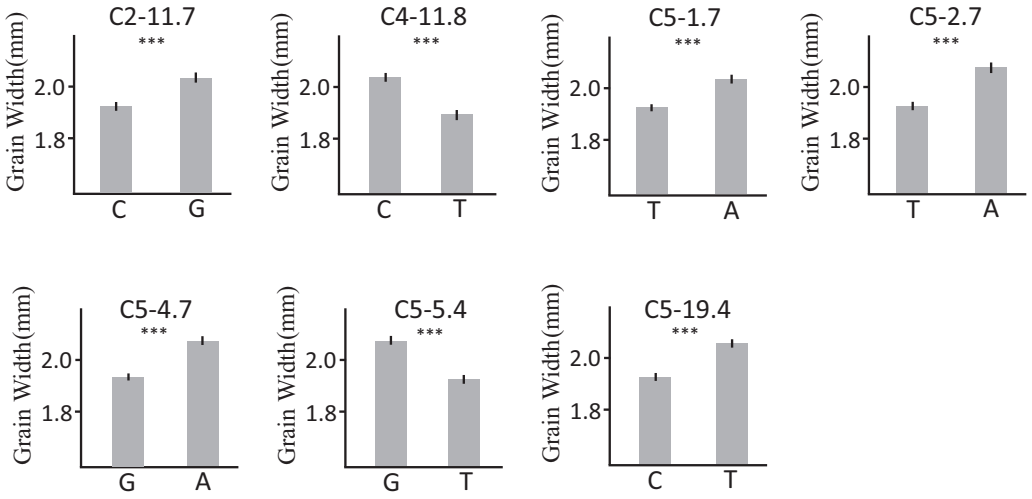

Table 2. Percentage accessions within rice subgroups that carry SNPs associated with increased grain width at each candidate locus.

Oryza sativa group	C2-11.7	C4-11.8	C5-1.7	C5-2.7	C5-4.7	C5-5.4	C5-19.4
subsp. *indica*							
indica	5.8	39.7	3.2	0	2.6	68.3	0
aus	0	100	0	0	0	4.5	0
subsp. *japonica*							
temperate japonica	96.4	98.8	88.2	94.1	100	90.6	90.6
tropical japonica	100	100	100	23.9	100	93.5	100
aromatic	100	100	90.0	0	20	100	80

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 predominating 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[2], S4C[2], S4G[2]). 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 japonicas* at C5-2.7; see Fig. S4D[2]). 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[2], S4F[2]). 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[2]). 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 θ_W. 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[2], S4E[2]). 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[2], S4F[2]). 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. Summary of nucleotide diversity and neutrality tests.

	C2-11.7			C4-11.8			C5-1.7			C5-2.7			C5-4.7			C5-5.4			C5-19.4		
Species/Group	π	θ_w	Tajima's D	π	θ_w	Tajima's D	π	θ_w	Tajima's D	π	θ_w	Tajima's D	π	θ_w	Tajima's D	π	θ_w	Tajima's D	π	θ_w	Tajima's D
Oryza sativa	0.00052	0.00019	3.4641**	0.00093	0.00041	3.55933***	0.00145	0.00084	3.24543***	0.00146	0.00080	1.67547	0.00182	0.00107	2.7753***	0.00219	0.00082	3.6334***	0.00054	0.00033	3.03392**
subsp. *indica*	0.00003	0.00021	−1.8759***	0.00061	0.00029	3.0901**	0.00008	0.00027	−0.9418	0.00126	0.00123	0.1411	0.00025	0.00073	−1.4111	0.00228	0.00090	3.5575***	0.00030	0.00040	−1.2534
indica	0.00001	0.00015	−1.2190	0.00055	0.00029	3.0631**	0.00004	0.00010	−1.5032	0.00072	0.00056	0.4599	0.00013	0.00097	−1.7794*	0.00227	0.00094	3.3686**	0.00046	0.00035	1.0212
aus	0.00003	0.00018	−1.7769	0.00069	0.00107	−0.0002	0.00006	0.00021	−1.1997	0.00065	0.00055	0.3725	0.00012	0.00085	−1.5212	0.00229	0.00115	2.7421**	0.00040	0.00036	1.1388
subsp. *japonica*	0.00019	0.00023	−0.3800	0.00062	0.00078	−0.6078	0.00134	0.00168	−0.5215	0.00202	0.00115	1.7637***	0.00137	0.00108	0.7031	0.00010	0.00050	−1.4852	0.00022	0.00036	−1.0454
aromatic	0.00015	0.00016	−0.2633	0.00041	0.00062	−1.1666	0.00051	0.00037	1.2348	0.00019	0.00027	−1.0062	0.00024	0.00022	0.5303	0.00044	0.00068	−1.3217	0.00015	0.00012	0.8504
tropical japonica	0.00016	0.00009	1.7579	0.00059	0.00073	−1.7072	0.00111	0.00125	−0.3326	0.00130	0.00073	1.8663	0.00025	0.00021	0.6384	0.00010	0.00085	1.9390***	0.00031	0.00025	0.3774
temperate japonica	0.00018	0.00022	−0.6248	0.00049	0.00070	−1.4105	0.00094	0.00109	−0.4053	0.00046	0.00065	−0.6494	0.00019	0.00016	0.7275	0.00009	0.00059	−1.8072*	0.00021	0.00040	−0.6631
O. rufipogon	0.00021	0.00002	0.6124	0.00238	0.00175	1.0895	0.00248	0.00145	1.66563***	0.00190	0.00175	0.2022	0.00239	0.00136	1.9489***	0.00240	0.00105	3.1924***	0.00072	0.00063	0.4013

Note: π, average number of nucleotide differences per site between two sequences (Nei 1987) calculated on the total number of polymorphic sites; θ_w, the Watterson estimator of θ per base pair (Watterson 1975) 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 SUPER 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 model 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 1; 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 (Pace 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[2]; Fig. S2[2]). 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[2]), 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 (*qSW5*), 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[2], S4F[2]). Similarly, Takano-Kai et al. (2009) documented that a mutation in the gene *GS3* 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 *GS5* and *GS6* QTLs were shown to be masked by *qSW5* (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 (2016YD0100301) 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 research as a visiting scholar in the Olsen Lab at Washington University. We thank members of the Olsen Lab group for helpful comments on the manuscript.

References

Aluko, G., Martinez, C., Tohme, J., Castano, C., Bergman, C., and Oard, J.H. 2004. QTL mapping of grain quality traits from the interspecific cross *Oryza sativa* × *O. glaberrima.* Theor. Appl. Genet. **109**(3): 630–639. doi:10.1007/s00122-004-1668-y. PMID:15105992.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T.L., et al. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. **3**(9): 1745–1756. doi:10.1371/journal.pgen.0030163. PMID:17907810.

Cao, Y.Y., Duan, H., Yang, L.N., Wang, Z.Q., Liu, L.J., and Yang, J.C. 2009. Effect of high temperature during heading and early filling on grain yield and physiological characteristics in *indica* rice. Acta Agron. Sin. **35**(3): 512–521. doi:10.1016/S1875-2780(08)60071-1. PMID:17907810.

Chapin, F.S.C., Autumn, K., and Pugnaire, F. 1993. Evolution of suites of traits in response to environmental stress. Am. Nat. **142**: 78–92. doi:10.1086/285524.

Cheng, W.G., Sakai, H., Yagi, H., and Hasegawa, T. 2009. Interactions of elevated [CO2] and night temperature on rice growth and yield. Agric. For. Meteorol. **149**(1): 51–58. doi:10.1016/j.agrformet.2008.07.006. PMID:22438302.

Cooper, N.T.W., Siebenmorgen, T.J., and Counce, P.A. 2008. Effects of nighttime temperature during kernel development on rice physicochemical properties. Cereal Chem. **85**(3): 276–282. doi:10.1094/CCHEM-85-3-0276.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. **14**(8): 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x. PMID:15969739.

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., et al. 2006. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor. Appl. Genet. **112**(6): 1164–1171. doi:10.1007/s00122-006-0218-1. PMID:16453132.

Fuller, D.Q. 2007. Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world. Ann. Bot. **100**(5): 903–924. doi:10.1093/aob/mcm048. PMID:17495986.

Fuller, D.Q. 2012. Pathways to Asian civilizations: Tracing the origins and spread of rice and rice cultures. Rice, **4**: 78–92. doi:10.1007/s12284-011-9078-7.

Fuller, D.Q., Qin, L., Zheng, X.M., Zhao, Z.J., Chen, X.G., Hosoya, L.A., and Sun, G.P. 2009. The domestication process and domestication rate in rice: spikelet bases from the lower Yangtze. Science, **323**(5921): 1607–1610. doi:10.1126/science.1166605. PMID:19299619.

Gnan, S., Priest, A., and Kover, P.X. 2014. The genetic basis of natural variation in seed size and seed number and their trade-off using *Arabidopsis thaliana* magic lines. Genetics, **198**(4): 1751–1758. doi:10.1534/genetics.114.170746. PMID:25313128.

Huang, X., and Han, B. 2014. Natural variations and genome-wide association studies in crop plants. Annu. Rev. Plant Biol. **65**: 531–551. doi:10.1146/annurev-arplant-050213-035715. PMID:24274033.

Huang, X.H., Wei, X.H., Sang, T., Zhao, Q.A., Feng, Q., Zhao, Y., et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. **42**(11): 961–976. doi:10.1038/ng.695. PMID:20972439.

Huang, X.H., Zhao, Y., Wei, X.H., Li, C.Y., 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**(1): 32–53. doi:10.1038/ng.1018. PMID:22138690.

Izawa, T., Konishi, S., Shomura, A., and Yano, M. 2009. DNA changes tell us about rice domestication. Curr. Opin. Plant Biol. **12**(2): 185–192. doi:10.1016/j.pbi.2009.01.004. PMID:19185529.

Kovach, M.J., Calingacion, M.N., Fitzgerald, M.A., and McCouch, S.R. 2009. The origin and evolution of fragrance in rice (*Oryza sativa* L.). Proc. Natl. Acad. Sci. U.S.A. **106**: 14444–14449. doi:10.1073/pnas.0904077106. PMID:19706531.

Leishman, M.R. 2001. Does the seed size/number trade-off model determine plant community structure? An assessment of the model mechanisms and their generality. Oikos, **93**(2): 294–302. doi:10.1034/j.1600-0706.2001.930212.x.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics, **25**(14): 1754–1760. doi:10.1093/bioinformatics/btp324. PMID:19451168.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. 2009. The sequence alignment/map format and samtools. Bioinformatics, **25**(16): 2078–2079. doi:10.1093/bioinformatics/btp352. PMID:19505943.

Li, L.F., Li, Y.L., Jia, Y., Caicedo, A.L., and Olsen, K.M. 2017. Signatures of adaptation in the weedy rice genome. Nat. Genet. **49**(5): 811–814. doi:10.1038/ng.3825. PMID:28369039.

Li, Y., Fan, C., Xing, Y., Jiang, Y., Luo, L., Sun, L., et al. 2011. Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. Nat. Genet. **43**(12): 1266–1269. doi:10.1038/ng.977. PMID:22019783.

Lu, L., Shao, D., Qiu, X., Sun, L., Yan, W., Zhou, X., et al. 2013. Natural variation and artificial selection in four genes determine grain shape in rice. New Phytol. **200**(4): 1269–1280. doi:10.1111/nph.12430. PMID:23952103.

Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., et al. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc. Natl. Acad. Sci. U.S.A. **110**: 453–458. doi:10.1073/pnas.1215985110. PMID:23267105.

Murray, M.G., and Thompson, W.F. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. **8**(19): 4321–4326. doi:10.1093/nar/8.19.4321. PMID:7433111.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., and Lübberstedt, T. 2015. Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genomics, **16**(1): 47. doi:10.1186/s12864-015-1226-9. PMID:25652714.

Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. **86**(6): 832–838. doi:10.1016/j.ajhg.2010.04.005. PMID:20471002.

Purugganan, M.D., and Fuller, D.Q. 2009. The nature of selection during plant domestication. Nature, **457**(7231): 843–848. doi:10.1038/nature07895. PMID:19212403.

Rozas, J. 2009. DNA sequence polymorphism analysis using DnaSP. Methods Mol. Biol. **537**: 337–350. doi:10.1007/978-1-59745-251-9_17. PMID:19378153.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., and Yano, M. 2008. Deletion in a gene associated with grain size increased yields during rice domestication. Nat. Genet. **40**(8): 1023–1028. doi:10.1038/ng.169. PMID:18604208.

Song, X.J., Huang, W., Shi, M., Zhu, M.Z., and Lin, H.X. 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nat. Genet. **39**(5): 623–630. doi:10.1038/ng2014. PMID:17417637.

Sun, L., Li, X., Fu, Y., Zhu, Z., Tan, L., Liu, F., et al. 2013. *GS6*, a member of the GRAS gene family, negatively regulates grain size in rice. J. Integr. Plant Biol. **55**(10): 938–949. doi:10.1111/jipb.12062. PMID:23650998.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics, **123**: 585–595. PMID:2513255.

Takano-Kai, N., Jiang, H., Kubo, T., Sweeney, M., Matsumoto, T., Kanamori, H., et al. 2009. Evolutionary history of *GS3*, a gene conferring grain length in rice. Genetics, **182**(4): 1323–1334. doi:10.1534/genetics.109.103002. PMID:19506305.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. **30**(12): 2725–2729. doi:10.1093/molbev/mst197. PMID:24132122.

Tang, Y., Liu, X.L., Wang, J.B., Li, M., Wang, Q.S., Tian, F., et al. 2016. GAPIT Version 2: An enhanced integrated tool for genomic association and prediction. Plant Genome, **9**(2): 1–9. doi:10.3835/plantgenome2015.11.0120. PMID:27898829.

Thomson, M.J., Tai, T.H., Mcclung, A.M., Lai, X.H., Hinga, M.E., Lobos, K.B., et al. 2003. Mapping quantitative trait loci for yield, yield components and morphological traits in an advanced backcross population between *Oryza rufipogon* and the *O. sativa* cultivar Jefferson. Theor. Appl. Genet. **107**(3): 479–493. doi:10.1007/s00122-003-1270-8. PMID:12736777.

Turner, D.S. 2014. qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots. Biorxiv. **1**: 1–2. doi:10.1101/005165.

Vaughan, D.A., Lu, B.R., and Tomooka, N. 2008. Was Asian rice (*Oryza sativa*) domesticated more than once? Rice, **1**(1): 16–24. doi:10.1007/s12284-008-9000-0.

Vigouroux, Y., Glaubitz, J.C., Matsuoka, Y., Goodman, M.M., Sanchez, G.J., and Doebley, J. 2008. Population structure and genetic diversity of new world maize races assessed by DNA microsatellites. Am. J. Bot. **95**(10): 1240–1253. doi:10.3732/ajb.0800097. PMID:21632329.

Wang, C.H., Zheng, X.M., Xu, Q., Yuan, X.P., Huang, L., Zhou, H.F., et al. 2014. Genetic diversity and classification of *Oryza sativa* with emphasis on Chinese rice germplasm. Heredity, **112**: 489–496. doi:10.1038/hdy.2013.130. PMID:24326293.

Wang, D., Xia, Y., Li, X., Hou, L., and Yu, J. 2013. The rice genome knowledgebase (rgkbase): an annotation database for rice comparative genomics and evolutionary biology. Nucleic Acids Res. **41**: 1199–1205. doi:10.1093/nar/gks1225. PMID:23193278.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., et al. 2014. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. **46**(9): 982–988. doi:10.1038/ng.3044. PMID:25064006.

Wang, Q., Tian, F., Pan, Y., Buckler, E.S., and Zhang, Z. 2014. A SUPER powerful method for genome wide association study. PLoS ONE, **9**(9): e107684. doi:10.1371/journal.pone.0107684. PMID:25247812.

Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., et al. 2012. Control of grain size, shape and quality by *OsSPL16* in rice. Nat. Genet. **44**(8): 950–954. doi:10.1038/ng.2327. PMID:22729225.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol, **7**: 256–276. doi:10.1016/0040-5809(75)90020-9. PMID:1145509.

Weng, J., Gu, S., Wan, X., Gao, H., Guo, T., Su, N., et al. 2008. Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. Cell Res. **18**(12): 1199–1209. doi:10.1038/cr.2008.307. PMID:19015668.

Westoby, M., Jurado, E., and Leishman, M. 1992. Comparative evolutionary ecology of seed size. Trends Ecol. Evol. **7**(11): 368–372. doi:10.1016/0169-5347(92)90006-W. PMID:21236070.

Xia, H.B., Xia, H., Ellstrand, N.C., Yang, C., and Lu, B.R. 2011. Rapid evolutionary divergence and ecotypic diversification of germination behavior in weedy rice populations. New Phytol. **191**(4): 1119–1127. doi:10.1111/j.1469-8137.2011.03766.x. PMID:21569036.

Xing, Y., and Zhang, Q. 2010. Genetic and molecular bases of rice yield. Annu. Rev. Plant Physiol. **61**: 421–442. doi:10.1146/annurev-arplant-042809-112209.

Xue, Y.D., Warburton, M.L., Sawkins, M., Zhang, X.H., Setter, T., Xu, Y.B., et al. 2013. Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. Theor. Appl. Genet. **126**: 2587–2596. doi:10.1007/s00122-013-2158-x. PMID:23884600.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.C., 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**(8): 927–936. doi:10.1038/ng.3596. PMID:27322545.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., 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**(2): 203–208. doi:10.1038/ng1702. PMID:16380716.

Zenni, R.D., Bailey, J.K., and Simberloff, D. 2014. Rapid evolution and range expansion of an invasive plant are driven by provenance–environment interactions. Ecol. Lett. **17**(6): 727–735. doi:10.1111/ele.12278. PMID:24703489.

Zhang, P., Zhong, K., Shahid, M.Q., and Tong, H. 2016. Association analysis in rice: From application to utilization. Front. Plant Sci. **7**: 1202. doi:10.3389/fpls.2016.01202. PMID:27582745.

Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat. Commun. **2**: 467. doi:10.1038/ncomms1467. PMID:21915109.

Zhou, W., Wang, Z., Davy, A.J., and Liu, G. 2013. Geographic variation and local adaptation in *Oryza rufipogon* across its climatic range in China. J. Ecol. **101**: 1498–1508. doi:10.1111/1365-2745.12143. PMID:26340227.