Title

The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication

Permalink
https://escholarship.org/uc/item/3582c9kz

Journal
Molecular Plant, 9(7)

ISSN
1674-2052

Authors
Wang, Hongru
Xu, Xun
Vieira, Filipe Garrett
et al.

Publication Date
2016-07-01

DOI
10.1016/j.molp.2016.04.018

Molecular Plant
Research Article

CellPress
PARTNER JOURNAL

The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication

Hongru Wang[1,2,7], Xun Xu[3,7], Filipe Garrett Vieira[4], Yunhua Xiao[1,2], Zhikang Li[5], Jun Wang[2,*], Rasmus Nielsen[6,*] and Chengcai Chu[1,*]

[1]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100101, China

[3]BGI-Shenzhen, Shenzhen 518083, China

[4]Centre for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark

[5]Institute of Crop Sciences/National Key Facilities for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, 12 South Zhong-Guan-Cun Street, Beijing 100081, China

[6]Department of Integrative Biology, University of California, Berkeley, CA 94720 USA

[7]These authors contributed equally to this article.

***Correspondence: Jun Wang (wangj@genomics.org.cn), Rasmus Nielsen (rasmus_nielsen@berkeley.edu), Chengcai Chu (ccchu@genetics.ac.cn)**

http://dx.doi.org/10.1016/j.molp.2016.04.018

ABSTRACT

Low-coverage whole-genome sequencing is an effective strategy for genome-wide association studies in humans, due to the availability of large reference panels for genotype imputation. However, it is unclear whether this strategy can be utilized in other species without reference panels. Using simulations, we show that this approach is even more relevant in inbred species such as rice (*Oryza sativa* L.), which are effectively haploid, allowing easy haplotype construction and imputation-based genotype calling, even without the availability of large reference panels. We sequenced 203 rice varieties with well-characterized phenotypes from the United States Department of Agriculture Rice Mini-Core Collection at an average depth of 1.5× and used the data for mapping three traits. For the first two traits, amylose content and seed length, our approach leads to direct identification of the previously identified causal SNPs in the major-effect loci. For the third trait, pericarp color, an important trait underwent selection during domestication, we identified a new major-effect locus. Although known loci can explain color variation in the varieties of two main subspecies of Asian domesticated rice, *japonica* and *indica*, the new locus identified is unique to another domesticated rice subgroup, *aus*, and together with existing loci, can fully explain the major variation in pericarp color in *aus*. Our discovery of a unique genetic basis of white pericarp in *aus* provides an example of convergent evolution during rice domestication and suggests that *aus* may have a domestication history independent of *japonica* and *indica*.

Key words: inbreeding, GWAS, rice, pericarp color

Wang H., Xu X., Vieira F.G., Xiao Y., Li Z., Wang J., Nielsen R., and Chu C. (2016). The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. Mol. Plant. **9**, 975–985.

INTRODUCTION

In humans, genome-wide association studies (GWAS) have successfully identified thousands of common genetic variants contributing to susceptibility to diseases (Hindorff et al., 2016). In inbred plant species, including many important crops such as rice and soybean, GWAS have the potential to be even more efficient for identifying phenotype–genotype associations because with just one-time genotyping of a population, the panel of inbred lines can be kept immortal in seed banks and can be phenotyped for different traits in different environments in both present and future studies. With the advances on next-generation sequencing (NGS) technology, DNA sequencing has become an appealing alternative to SNP arrays for genotyping, extending GWAS beyond common variants, and holding the promise to capture rare alleles and structural variants. In humans, even with extremely low sequencing depth

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

(0.1–0.5×), NGS-based GWAS have statistical power commensurate to that of SNP array-based GWAS (Pasaniuc et al., 2012). The critical component for retaining high statistical power at such a low coverage is genotype imputation relying on the availability of a comprehensive panel of reference haplotypes. However, such reference panels of haplotypes are not available for most other species.

In rice, both SNP array (Zhao et al., 2011) and sequencing-based genotyping strategies (Huang et al., 2010, 2012b; Chen et al., 2014) have been adopted for GWAS covering diverse sets of traits. These efforts have systematically examined the genetic architecture of agronomic traits (Huang et al., 2010, 2012b), metabolism (Chen et al., 2014), and gene–environment interactions (Zhao et al., 2011) in rice, and have contributed a wealth of genomic and germplasm resources for both the research community and rice breeders. Genome sequencing-based strategies for GWAS have also found wide applications in other crops, including foxtail millet (Jia et al., 2013), sorghum (Morris et al., 2013), maize (Li et al., 2013; Wen et al., 2014), soybean (Zhou et al., 2015), and sesame (Wei et al., 2015). However, no study has systematically examined the impact of sequencing depth on the mapping power of GWAS in the crop populations with inbred genomes. In this study, we address this question by exploring the potential of low-coverage sequencing strategy for GWAS in inbred species. Even in the absence of large reference panels, low-coverage sequencing may be highly effective in inbred species that are effectively haploid and thereby avoid some of the technical challenges associated with diploid genotype calling. We show that in inbreeding species, GWAS using low-coverage (1–2×) sequencing data has similar power as that using genotypes from high-coverage data even when utilizing a *de novo* SNP discovery strategy. Furthermore, we sequence 203 inbred lines in the United States Department of Agriculture (USDA) Rice Mini-Core Collection (Agrama et al., 2009), at an average depth of 1.5×. This collection contains a representative subset of the rice entries in the USDA rice germplasm collection (Agrama et al., 2009; Li et al., 2010), making it suitable for mapping studies aimed at detecting genetic variation segregated in a diverse set of rice varieties. Using this strategy, we show that two agronomically important loci can be mapped to the level of causal SNP. We also discover a new locus contributing to pericarp color missed by previous mapping studies.

RESULTS

Sequencing Depth and GWAS Mapping Power in Inbreeding Species

To evaluate the power of GWAS by low-coverage sequencing data in inbred species, we used coalescent-based simulations to generate samples consisting of fully inbred individuals with sequencing data at various depths. We adopt a model mimicking rice demography and biology (see Methods) so that the patterns of mutation and local recombination can be accounted for realistically. We first evaluate false and true SNP discovery rate and genotype calling accuracy, then evaluate mapping power under different sequencing depth and sample size scenarios.

As expected (Li et al., 2011b; Fumagalli, 2013), SNP discovery rate is largely decided by the joint effect of sample size and sequencing depth (Supplemental Figure 1). The probability of detecting an SNP with allele frequency, f, in a fully genotyped sample of n haploid individuals is

$$h_n(f) = 1 - f^n - (1 - f)^n \ . \qquad \text{(Equation 1)}$$

Assuming the classical distribution of allele frequencies proportional to 1/f, the detection probability of SNPs with a minor allele frequency (MAF) greater than 5% in the population is then

$$\frac{\int_{0.05}^{0.95} h_n(f)/f df}{\int_{0.05}^{0.95} 1/f df} \ . \qquad \text{(Equation 2)}$$

This upper bound, corresponding to fully known genotypes, is effectively achieved at 5× sequencing depth in different sample size scenarios, using an ascertainment criterion of a likelihood ratio >15 (Supplemental Figure 2). In the small sample size scenario, e.g. 30 individuals, the detection power drops dramatically with sequencing depth. For example, 96% of polymorphic sites with MAF >5% can be detected at 5× sequencing depth, but only 87% at 1×, 77% at 0.5×, and 31% at 0.1×. However, the detection power loss can be compensated by increasing sample size. With a sample size of 200 or greater, sequencing at 0.5× can achieve detection power essentially identical to that of true genotypes. The false-positive rate in SNP discovery, using the likelihood ratio SNP ascertainment criterion, is low overall, and drops dramatically as sample size and sequencing depth increases. For the 1× dataset, the false-positive rate is essentially 0 when the sample size is 100 or larger (Supplemental Figure 2).

The genomes of inbred species are effectively haploid, greatly reducing the genotype calling problems associated with diploid data. Many studies have used diploid genotype callers, designed for outbred species, when calling genotypes in inbred genomes (Huang et al., 2010, 2012b; Jia et al., 2013; Li et al., 2013; Morris et al., 2013; Chen et al., 2014; Wen et al., 2014; Wei et al., 2015; Zhou et al., 2015), thereby in reality not taking advantage of the reduced complexity of the problem. However, methods for genotype calling from populations with different level of inbreeding are available (Vieira et al., 2013). By incorporating the inbreeding coefficient of each individual as prior when calling genotype, genotype calling in inbred species can be greatly improved. In our simulations with 200 fully inbred individuals, the genotype calling accuracy is close to 100% for all sequencing depth scenarios when applying stringent genotype filtering (see Methods and Supplemental Figure 3). Using high-quality genotypes, we carried out genotype imputation for the remaining SNPs with Beagle (Browning and Browning, 2007). Imputation based on linkage disequilibrium (LD) is effective in inbreeding species, because haplotypes are directly available in the inbred regions and LD also decays more slowly. For the 0.1× dataset, in which 91% of genotypes are missing data after initial genotype calling, the accuracy of imputed genotypes is poor at 76%, but quickly climbs to 94% at a sequencing depth of 0.5×, under which 48% of genotypes are missing. For 1× data or above, all missing data could be imputed without significant loss of accuracy (Supplemental Figure 3).

With the fully imputed genotype datasets, we calculated mapping power as a function of sequencing depth and h^2, the

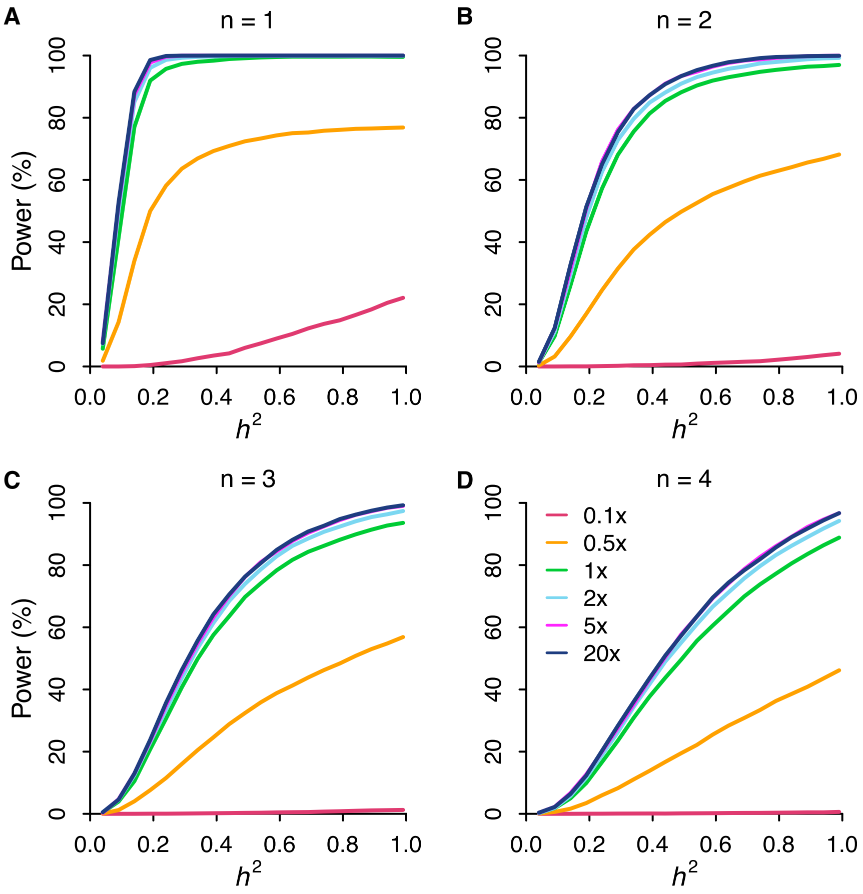

Figure 1. Mapping Power by Sequencing Depth, Narrow-Sense Heritability, and Number of Causal Loci.

(A–D) We simulated phenotypes for different numbers of causal SNPs (*n* ranges from 1 to 4 in **A–D**) and narrow-sense heritability (h^2), for 200 inbred individuals at depths of 0.1×, 0.5×, 1×, 2×, 5×, and 20×. The mapping power was calculated as the percentage of tests where the causal SNPs (or flanking SNPs) were associated with the trait with genome-wide significance at the 0.05 significance level.

narrow-sense heritability of a trait, which can also be interpreted as the proportion of total phenotypic variance caused by all causal SNPs combined for a polygenic trait (Figure 1). In all scenarios, 1× data provide similar power to that of 20× data in terms of both mapping power (Figure 1) and power to identify causal variants (Supplemental Figure 4), indicating that there is essentially no statistical advantage in terms of mapping power in using high-coverage sequencing data. At a sequencing depth of 0.5×, the mapping power is only slightly reduced for a Mendelian locus (n = 1) but more dramatically reduced in other cases. At h^2 = 0.59, the 0.5× sequencing scenario provides maximal power of 93% for a Mendelian locus. When phenotypic variation is contributed by multiple quantitative trait loci (QTLs), the 0.5× scenario has reduced power compared with the 20× scenario. If the sequencing depth is extremely low (0.1×), the mapping power is greatly reduced, never reaching 50% of that of 20× datasets in any scenario. The loss of power is largely due to the loss in the power of SNP discovery and the reduction in effective sample size due to imperfect imputation. Therefore, based on the simulation results, we recommend a sequencing depth of 1–2× when designing sequencing-based GWAS in inbred species.

Genome Sequencing of a Diverse Rice Collection

We sequenced the USDA Rice Mini-Core Collection (Agrama et al., 2009), a rice population consisting of 203 diverse rice germplasm accessions (Supplemental Table 1). This collection was systematically developed from 1794 core entries in the USDA rice collection based on both phenotypic and genotypic data, and is considered to be a representative subset of more than 18 000 accessions of rice entries worldwide in the USDA rice germplasm collection (Figure 2A) (Agrama et al., 2009; Li et al., 2010). The samples are from 75 countries spanning the regions of genetic diversity and cultivation of rice (Figure 2B and Supplemental Figure 5). In total, 1280 million reads were generated, representing an average genomic coverage of 1.5× for each accession. We employed a short-read remapping alignment strategy with Stampy (Lunter and Goodson, 2011) that led to significant improvement in mapping quality when compared with the strategy using BWA (Li and Durbin, 2009) only (Supplemental Figure 6). The remapping strategy was especially successful in *indica* rice, with a 10% increase in the ratio of mapped paired end reads with mapping quality score >30. Overall, 65% of genomic regions of each accession were covered by at least one high-quality sequencing read (Supplemental Figure 7). After mapping the sequenced reads of each individual to the rice genome, we called SNPs in non-repetitive genomic regions of this population with ANGSD (Korneliussen et al., 2014). In total, 2.3 million high-confidence SNPs were obtained, representing six SNPs per kilobase pair.

We adopted genotype likelihoods-based methods to analyze the genetic structure of this population (Fumagalli et al., 2014). These methods account for the uncertainty regarding genotypes inherent in low-coverage data (Nielsen et al., 2011). NGSadmix analyses split the mini-core population into five ancestries corresponding to *indica*, *aus*, *aromatic*, *temperate japonica*, and *tropical japonica* under the K = 5 model (Figure 2). Principal component analysis (PCA) (Supplemental Figure 8) decomposed the population into five clusters corresponding to the five subgroups. Notably, there are large numbers of admixed accessions located in the PCA space between major clusters, reflecting a complicated history of hybridization and differentiation in rice cultivation history. The first four PCs add up to explain 21.6% of the total SNP variation, which is much less than in previous studies (Zhao et al., 2010, 2011). This is largely due to the fact that genome sequencing in this study captures many more rare alleles than previous studies based on SNP arrays.

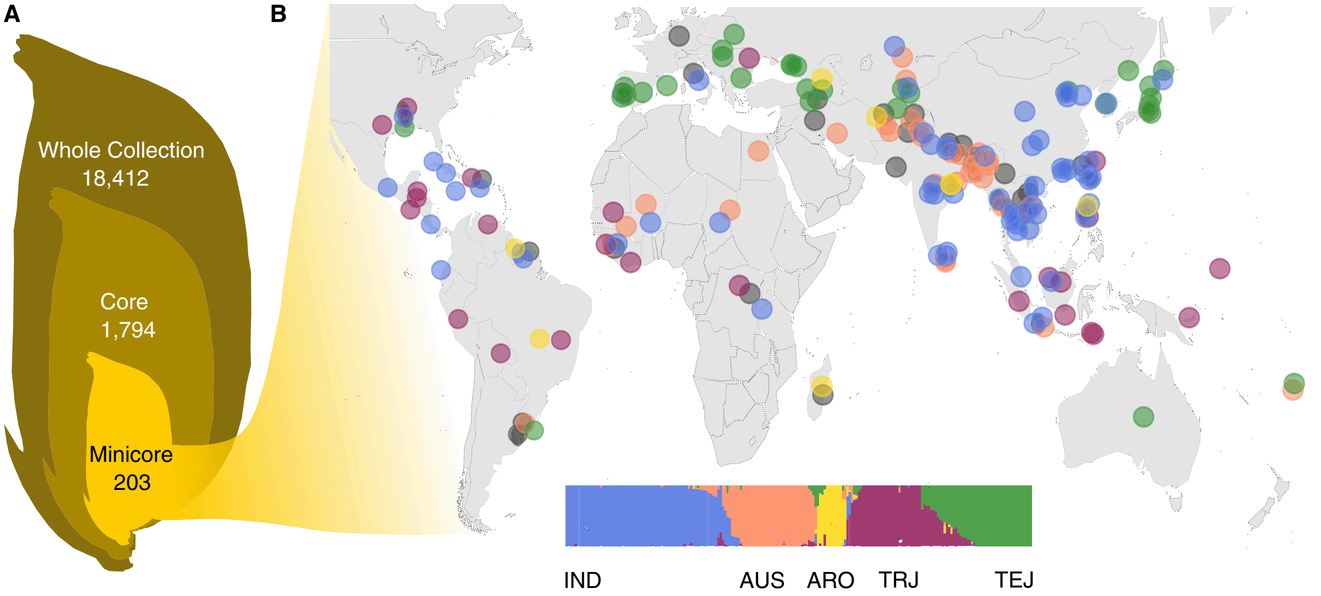

Figure 2. The USDA Rice Mini-Core Collection.

(A) The Mini-Core Collection was developed to retain the genetic diversity of 1794 core entries, which again was selected to be representative entries from more than 18 000 accessions in the USDA whole rice germplasm collection.

(B) The 203 domesticated entries in the Mini-Core Collection comes from 75 countries. Each dot on the world map represents one variety. The horizontal bar below summarizes the distribution of subpopulations of all domesticated entries in the Mini-Core Collection. The ancestry of each entry was inferred using NGSadmix. The color codes are blue for *indica* (IND), orange for *aus* (AUS), gold for *aromatic* (ARO), deep pink for *tropical japonica* (TRJ) and green for *temperate japonica* (TEJ).

Domesticated rice is a mostly self-pollinating plant, and rice germplasm has been held in seed banks for many generations through selfing. Therefore, their genomes are expected to be mostly homozygous. In accordance with this expectation, we found that most of the rice accessions have estimates of inbreeding coefficients (*F*) close to 1 (Supplemental Figure 9). Using a prior for genotype calling based on these estimates of *F* (Vieira et al., 2013), we called genotypes at SNP sites for each individual and obtained a genotype dataset with an overall missing rate of 32.8%. We evaluated the accuracy of the genotype dataset by Sanger sequencing. The concordance rate between the genotypes obtained from the two experiments is 98.9%. After imputation, the missing data were fully inferred, but the concordance rate remained at 98.5% (Supplemental Table 2).

GWAS on Grain Length

Previous studies have accumulated comprehensive phenotypic data on the Mini-Core Collection (Yan et al., 2009; Li et al., 2010, 2011a, 2012; Bryant et al., 2011; Jia et al., 2012). With the whole-genome SNP markers now available, we can conduct GWAS with unprecedented resolution for this mapping population. We here present results for seed length, amylose content, and pericarp color. GWAS was carried out using GEMMA (Zhou and Stephens, 2012). We notice that the use of a relatedness matrix as covariate in GEMMA corrects for the confounding factor of population structure (Figure 3). Genome-wide critical values were obtained using permutation (see Methods).

For grain length, only one signal exceeded the 5% genome-wide critical value (Figure 3A and 3B). The significant SNP is located in the genic region of *GS3* (Os03g0407400), a gene underlying a major QTL controlling the grain length and weight of rice (Fan et al., 2006). In fact, the significant SNP is a C/A polymorphism in the second exon of the gene, which introduces a stop codon and truncates the amino acid sequence at position 178 of the expressed protein. Previous studies have also identified this SNP as strongly associated with grain length (Fan et al., 2006) and have identified it as a causal mutation (Mao et al., 2010). We detected two other SNPs in our rice population (Chr3:16731513 and Chr3:16732415) in this gene, neither of which, however, correlated with the phenotype ($P = 0.90$ and $P = 0.04$, respectively), adding further evidence to the previous studies arguing that the C→A mutation is causal.

GWAS on Amylose Content

We then performed a GWAS on amylose content in rice kernels, a major determinant of rice stickiness, and found a major peak on chromosome 6 (Figure 3C and 3D). This region harbors 121 SNPs highly associated with the trait. The peak SNP, Chr6:1765761, is a G/T polymorphism located on the genic region of *Wx* (Os06g0133000), which encodes a starch synthase (Supplemental Figure 10). The gene is the major QTL for amylose content in rice endosperm (Wang et al., 1995). Previous studies have showed that the Chr6:1765761 SNP falls on splice site 5 of *Wx*, with the T allele leading to dramatically reduced levels of spliced mRNA by post-transcriptional regulation, giving rise to glutinous rice (Wang et al., 1995; Hirano et al., 1998). In our rice population, rice accessions with T and G alleles have average amylose content of 13.7% and 24.1%, respectively (Supplemental Figure 11). These observations support the conclusion of previous studies (Wang et al., 1995;

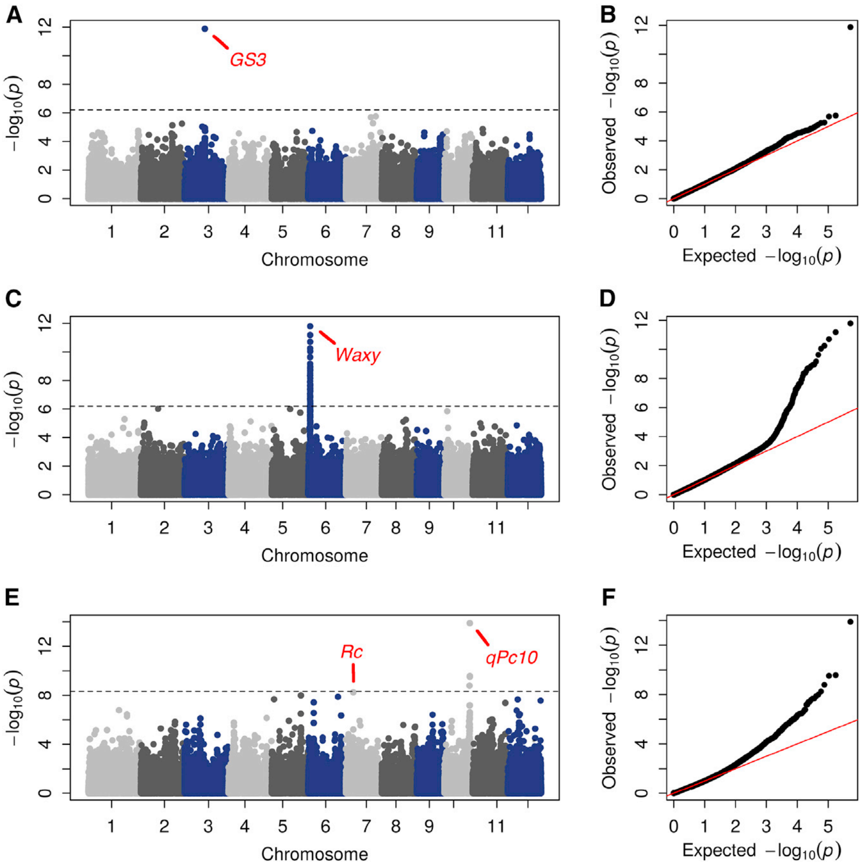

Figure 3. GWAS for Agronomic Traits in Rice.

(A, C, and E) Manhattan plot of genome-wide association results for **(A)** grain width, **(C)** amylose content, and **(E)** pericarp color in rice. The dotted horizontal line marks a significance level of 0.05 after correction for multiple tests as determined by independent permutations for each trait.

(B, D, and F) The corresponding QQ plot of the distribution of observed *P* values versus those expected under the null for the three GWAS.

Hirano et al., 1998) that SNP Chr6:1765761 is the causal polymorphism for amylose content variation. Chr6:1765761 is located in the center of an LD block, which is approximately 250 kb long (Supplemental Figure 12), suggesting that strong artificial selection on the gene has given rise to a selective sweep in this region (Olsen et al., 2006).

In the cases of both amylose content and seed length, our study design had power to identify major QTLs of agronomically important traits. Furthermore, the high-density marker set provides power to identify the causal SNP with high probability. However, these traits had been extensively studied previously, and our study could only confirm the conclusions of previous studies. In the next example, we illustrate the use of the design to identify novel QTLs for rice pericarp color.

GWAS on Pericarp Color

Rice pericarp coloration differs between wild and cultivated rice, with the vast majority of wild or weedy species of *Oryza* having red pericarp color, while most cultivars have a white pericarp (Sweeney et al., 2007). The red pericarp of rice has attracted research interest for a century because of its convenient use as a genetic marker (Kato and Ishikawa, 1921), and more recently because of its implications in rice domestication (Izawa et al., 2009) and also its potential nutritional importance (Shirley, 1998). The red pigmentation in rice grains is proanthocyanidin, whose biosynthetic pathway is mostly shared with that of anthocyanidin (Furukawa et al., 2007). Previous molecular studies identified two genes, *Rc* and *Rd*, to be necessary for red pigmentation in rice grains (Sweeney et al., 2006; Furukawa et al., 2007). Sweeney et al. (2006) further identified two independent loss-of-function alleles of *Rc*: *rc*, characterized by a 14-bp deletion in *Rc*, and *Rc-s*, characterized by a C → A mutation resulting in the truncation of the protein before the helix-loop-helix domain. The *Rcrd* genotype produces brown pigmentation and *RcRd* red pigmentation in the pericarp (Sweeney et al., 2006; Furukawa et al., 2007). Rice harboring *rc* allele has a white pericarp, and this is the common allele shared by white-pericarp rice of both *indica* and *japonica* varieties (Sweeney et al., 2007). *Rc-s* allele is specific to *aus* varieties which, for unknown reasons, produces both white and light-red pericarps, independently of the alleles in the *Rd* (Sweeney et al., 2007).

We conducted a GWAS on rice pericarp coloration (Figure 3E and 3F) and we first examined the polymorphisms in *a priori* defined gene regions. In the genic region of *Rd*, 15 SNPs were detected from the population, among which there are two SNPs known to affect function, *Rd1* and *Rd2* (Konishi et al., 2008), which introduce two independent stop codons. However, we found none of the SNPs to be significantly associated with the phenotype. This is likely due to the fact that *Rd* only helps to increase pigment content in the presence of the *Rc* allele, and the mutations in *Rd* do not eliminate pericarp coloration (Sweeney et al., 2006; Furukawa et al., 2007; Konishi et al., 2008). In the *Rc* gene region, the causal *Rc-s* SNP at Chr7:6068017 was found to be strongly associated with pericarp coloration ($-log_{10}(P)$ = 8.1), representing the second highest peak in the Manhattan plot (Figure 3D). Examination on the associated SNP in our study showed that the signal is exclusively driven by rice from the *aus* and *admixture* subgroups (Supplemental Table 3). Since the 14-bp mutation in the *rc* allele is not included in our genotype dataset, and it is likely that no SNP linked with the mutation was included either, we find no polymorphism in *indica*/*japonica* varieties that are highly associated with the phenotype in *Rc* region. Intriguingly, there are four *aus* varieties with genotype–phenotype associations that are not predicted by the results of previous studies. GSOR310703 and GSOR311111 have the C allele at the Chr7:6068017 position, but their pericarps are white. We thus predicted that these

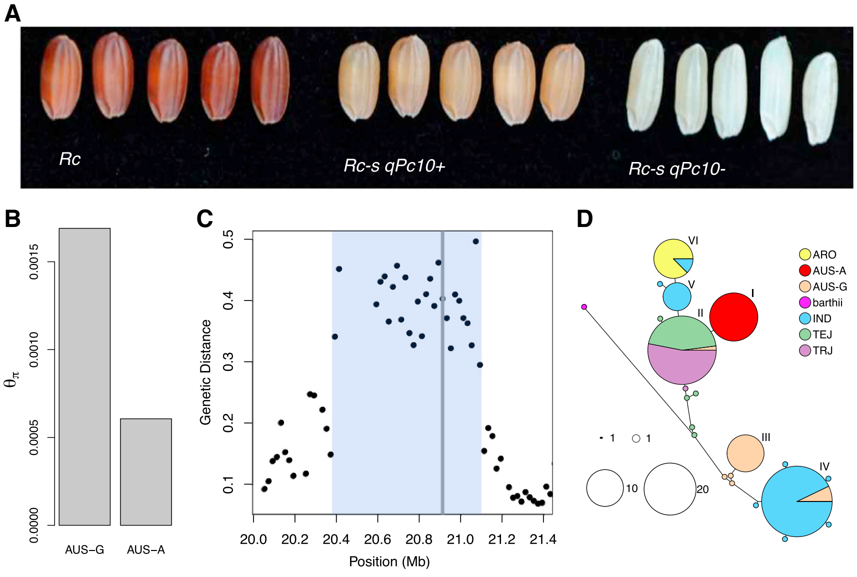

Figure 4. Evolution of Pericarp Color in *aus*.

(A) Pericarp color variation in *aus* varieties. *Aus* harboring *Rc* or *rc* has red pericarp. When containing an *Rc-s* allele, the pericarp color *aus* varieties is light red in the context of *qPc10*$^+$ and white in the context of *qPc10*$^-$.

(B) θ_π estimated in both AUS-A and AUS-G populations.

(C) Genetic distance between AUS-A (*aus* varieties harboring allele A at Chr10:20912658) and AUS-G is highly elevated in the *qPc10* region. Each dot represents the average value of genetic distance calculated over all polymorphic sites in 20-kb region. The x axis is the physical position along chromosome 10. The blue box defines a ~700 kb region with increased genetic differentiation between AUS-A and AUS-G.

(D) A haplotype network based on the number of pairwise differences among haplotypes in a 20-kb region harboring Chr10:20912658 in the Mini-Core population. We used pegas (Paradis, 2010), an R package, to generate the figure. Each pie chart represents one unique haplotype, and the radius of the pie chart is proportional to the log$_2$(number of chromosomes with that haplotype) plus a minimum value. Six major haplotype groups were labeled with Roman numerals on the upper right. The lengths of edges are proportional to the number of differences, and the scale is shown at the button left. ARO, *aromatic*; barthii, *Oryza barthii*; IND, *indica*; TEJ, *temperate japonica*; TRJ, *tropical japonica*.

varieties harbor the *rc* allele. Examination on reads mapped to this region confirmed our prediction with four and two high-quality reads (mapping quality >30) support, respectively. Two other varieties, GSOR310945 and GSOR311181, harbor *Rc-s* but have light-red pericarp. Thus, the *Rc* locus alone is inadequate for explaining the white pericarp of *aus* varieties. In our GWAS, another association signal was observed on Chr10, representing a new QTL for rice pericarp coloration (henceforth denoted *qPc10*). The peak SNP, located at Chr10:20912658, is a G/A polymorphism with $-log_{10}(P)$ of 13.9, representing the strongest association with phenotype among all SNPs in the sample. Interestingly, the statistical power at this site is also mostly contributed by *aus/admixture* varieties (Supplemental Table 4).

To validate the association signals, we scored the pericarp coloration for 206 *aus* varieties from the 3000 rice genomes project (Alexandrov et al., 2015; The 3000 Rice Genomes Project, 2014). Association tests showed that both signal SNPs discovered in our GWAS are significantly correlated with pericarp coloration in this independent worldwide *aus* varieties panel ($P = 2.2 \times 10^{-16}$ for Chr7:6068017 and $P = 1.6 \times 10^{-11}$ for Chr10:20912658, Spearman's ρ test). Interestingly, the two rice accessions in the Mini-Core Collection that contain the *Rc-s* allele but have light-red pericarp also have the ancestral allele (G) at Chr10:20912658, and all *aus* varieties with white pericarp harbor the mutated allele (A) at the site. Thus, *qPc10* is the missing locus, which, combined with the *Rc* locus, can fully explain the pericarp coloration in *aus* varieties (Figure 4A and Supplemental Table 5). The Chr10:20912658 SNP is located in an intergenic region. To find candidate genes for *qPc10*, we searched the flanking regions. One possible candidate gene is *Os10g0536400* located 23 kb upstream of Chr10:20912658, which codes for a flavanone 3-hydroxylase (F3H). F3H catalyzes the chemical reaction from flavanones to dihydroflavonols, which is the first committed step of the biosynthesis of anthocyanins and proanthocyanidins, the two major compounds responsible for the coloration of rice pericarp (Furukawa et al., 2007). Through PCR amplification and sequencing, we identified 151 variants on the gene and 2-kb promoter region of *Os10g0536400* in the *aus* population, but none of the variants generate any association signal stronger than Chr10:20912658. It is possible that differences in expression level cause the phenotypic variation. However, further genetic and molecular evidence, including examination of the expression profile of the gene on the same genomic background, are needed to validate the involvement of this gene in rice pericarp coloration.

Evolution of Pericarp Color in *aus*

The accumulation of proanthocyanidin in grains of wild rice contributes to plant defense against pathogens or predators. Mutations leading to white pericarp are therefore thought to be selected against in nature (Shirley, 1998). However, the pericarp of most cultivated rice is white, suggesting that selection for white pericarp was induced during domestication by farmers (Sweeney et al., 2007). Our findings show that the genetic basis of pericarp color in *aus* is different from that in *indica* and *japonica* rice, suggesting independent selection on this domestication trait in *aus*. The first step in the evolution toward white pericarp in *aus* involved an independent stop-codon mutation in *Rc* (*Rc-s*), which now has a frequency of 86.7% (13/15) in white-pericarp *aus* varieties in our rice panel and 91.5% (43/47) in white-pericarp *aus* varieties from the 3000 rice genomes project (Alexandrov et al., 2015). This mutation turned the pericarp color of rice from red to light red. Next, when a mutation in *qPc10* was introduced onto the *Rc-s* background, white pericarp in *aus* arose. Consistent with this scenario, we detected a reduction of nucleotide diversity in *qPc10* region by 64.0% in *aus* accessions with the derived allele (AUS-A) at Chr10:20912658 compared with *aus* accessions with the ancestral allele (AUS-G) at the site (Figure 4B). The extent of diversity in the region is significantly reduced among haplotypes carrying the derived allele ($P = 0.017$, Supplemental Figure 13). The genetic distance between AUS-A and AUS-G, which was measured as an average distance between two sequences that are randomly drawn from two populations, is 0.40. This is significantly elevated when compared with the genomic background value of 0.06 (P = 0.0044, Supplemental Figure 14). These results support the hypothesis of a recent selective sweep that eliminated haplotype diversity in this region in the *aus* population, likely due to selection imposed by farmers in favor of white rice. When we extend the genetic distance calculation to the flanking regions, we find a highly differentiated block with sharp boundaries spanning ~700 kb from position 20 400 000 to 21 100 000 (Figure 4C). This is remarkable because it suggests that AUS-A and AUS-G form haplotype groups that are divergent from each other in this large genomic region. This pattern is not easy to explain just by a scenario of selection from standing variation, but may suggest that the haplotype was introduced by gene flow from a divergent variety and/or has been experiencing recombination suppression. To test the gene-flow hypothesis, we searched for the source of both haplotype groups from other rice subgroups in our sequenced rice panel and constructed a haplotype network including all haplotypes from five major rice subgroups using all SNPs within a 20-kb region (Figure 4D). The network consists of two major clusters, one containing most *indicas* and one of the two major differentiated haplotype groups within *aus* or haplotype III (AUS-G), and the other containing all *japonicas*, *aromatic* varieties, some *indicas* plus another *aus* subgroup or haplotype I (AUS-A) (Figure 4D). The minimal distance between haplotype groups I and III is 82 nucleotide differences. This pattern suggests that AUS-A varieties have obtained this haplotype by gene flow from *japonica* varieties. Moreover, haplotype I is slightly more divergent from haplotype III than from haplotype II (largely consisting of *japonica* varieties). This pattern is consistent with expectations under the hypothesis that AUS-A was introduced into *aus* from *japonica* varieties.

DISCUSSION

Low-coverage sequencing-based GWAS in inbreeding species using a *de novo* SNP discovery strategy is a cost-effective approach for association mapping. With both simulated and real data, we have shown that a sequencing depth of approximately 1× provides almost the same mapping power for GWAS as deep sequencing. In fact, in the real data, even with reasonably small panels, we were in several cases able to identify the causative SNP directly. The critical component for this framework is the use of population panels and imputation-based methods for genotype calling. The lesson from human genetics, that low-coverage sequencing can be highly effective (Pasaniuc et al., 2012), is even more true for inbreeding species which, for the purpose of genotype calling, can be considered to be haploid. This leads to easy resolution of haplotypes and much higher genotype calling accuracy. This framework does not rely on existing comprehensive haplotype panels, and thus should have an immediate application to GWAS in many other inbreeding species where such resources are not available. For computational reasons, our simulations used an approach in which we first called genotypes without imputation and then imputed the missing genotypes. For real data analyses we would instead recommend using direct imputation-based genotype calling (e.g., Browning and Browning, 2007). Such methods will likely greatly improve the performances of low-coverage approaches (e.g., 0.1×) over that reported here. However, for low-coverage whole-genome resequencing projects, the limiting factor for further reducing cost has shifted to the cost of sample preparation (Rohland and Reich, 2012).

When we applied the approach to real data, we identified a new locus for pericarp color variation in *aus* rice. Interestingly, white pericarp seems to have evolved twice, providing an example of convergent evolution in rice domestication. This result suggests that much of the domestication history of *aus* is independent of that of *japonica* and *indica*. Either *aus* represents an independently domesticated rice type that subsequent to early domestication received extensive gene flow from other rice varieties, or it arose from the same early domestication event but then split up from other domesticated varieties before white rice had evolved in these varieties.

The Rice Mini-Core Collection sequenced in the current study contains immortalized germplasm resources and is available through the seed bank (http://www.ars.usda.gov/Main, accessed on March 7, 2016). The enormous genetic diversity of rice represented in this collection has attracted researchers to join the efforts for phenotypic characterization, and as a result comprehensive phenotypic information, including field and kernel traits (Li et al., 2010, 2011a, 2012; Bryant et al., 2011), disease resistance (Jia et al., 2012), and rice grain nutrients (Pinson et al., 2015), have been accumulated. The genomes sequenced in this study provide a new opportunity for interrogating the genetic basis underlying these traits. As illustrated here, the diversity of the collection, in combination with low-coverage sequencing, provides sufficient power to map major-effect loci, even to the level of individual causal SNPs. In addition, it provides a basis for new experimental approaches for studying phenotypes experimentally under different environments, providing an accessible open-source platform for synthetic studies of genotype–phenotype associations and gene–environment interactions in rice.

METHODS

Simulation of Sequencing Data

To realistically simulate populations in MSMS (Ewing and Hermisson, 2010), we used population parameters specific to rice. We assumed an effective population size of 125 000 (Caicedo et al., 2007), a mutation rate of 6.5×10^{-9} (Gaut et al., 1996), and a recombination rate of 4 cM/Mb (Tian et al., 2009). We simulated 10 000 DNA fragments of 2 Mb, representing a population consisting of 10 000 fully inbred diploid individuals. The command line used for MSMS was: ‘‘java -jar msms.jar 10000 1000 -t 10000 -r 40000 2000000.’’ The population simulation was replicated 1000 times for downstream analysis. From each simulated population, we randomly drew samples of 30, 60, 100, 200, and 400 individuals. For each individual in a sample, we used ART (Huang et al., 2012a), a software that simulates the Illumina sequencing process with an empirical error model, to generate 100-bp non-gapped, mapped reads in SAM format of various depth directly (‘‘-len 100 -ir 0 -dr 0 -ir2 0 -dr2 0 –sam’’). We simulated sequencing data with depth ranging from 0.1× to 20× for each individual.

SNP Detection and Error Estimation for Simulations

The package ANGSD (Korneliussen et al., 2014) (version 0.542) was used for SNP detection. The implemented algorithm calls SNPs across individuals based on whether the minor allele frequency is significantly larger than 0 as determined by a likelihood ratio test (Nielsen et al., 2011). We used the "minLRT $\geq$15" criterion, which corresponded to a significance level of approximately 0.0001 for rejecting the hypothesis of the site being non-polymorphic, thereby inferring an SNP. To reduce potential false positives introduced by mismapping at gapped region, we included the arguments "-baq 1 -C 50." The arguments "-minMapQ 20 -minQ 20" were also used to restrict genotype calling to sites covered by reads with both high mapping and base quality. For genotype calling, we adopted a population-based strategy but corrected for the inappropriate Hardy–Weinberg equilibrium assumption by using the inbreeding coefficient of each individual as prior. To further control the genotype quality, we restricted genotype calling to sites covered by at least one high-quality read ("-geno_minDepth 1"). We also set the threshold for genotype posteriors to be greater than 0.9 ("-postCutoff 0.9").

The filtered SNP dataset was compared with the simulated true genotypes in sites with MAF $\geq$5%, to estimate the power of SNP discovery, SNP false discovery rate, and genotype accuracy under different sample size and sequencing depth scenarios. The SNP discovery power was calculated as the proportion of SNPs with MAF $\geq$5% in the population that were detected to be polymorphic. The false discovery rate was calculated as the proportion of SNPs that were called as polymorphic but actually were not polymorphic in the population. The genotyping accuracy was measured as the proportion of correct genotype calls.

Phenotype Simulation

We assumed a simple model of additive and equal effects (e) among loci. The phenotype of each individual was simulated from an $N(ne, V_E)$ distribution, where V_E is the phenotypic variance explained by environmental factors and n is the number of loci in which the individual carries a causal mutation. Phenotypes were simulated under different numbers of causal loci (n, ranging from 1 to 4) and different narrow-sense heritability (h^2, ranging from 5% to 100% with a step of 5%). For each combination of h^2 and n, we first randomly drew n sites from the genotype dataset with MAF $\geq$5% of the population as causal loci. For each chosen locus, we randomly let one of the alleles be causal and recorded the corresponding allele frequency f. The effect of each locus (e) can thus be calculated by solving the following equations: $h^2 = \frac{V_G}{V_E + V_G}$ and $V_G = \sum_{i=1}^{n} e^2 f_i (1 - f_i)$. Without loss of generality, we let $V_E + V_G = 1$. For each individual, n can then be tabulated.

GWAS

We used imputed genotype datasets for GWAS. Genotype imputation was performed on genotypes obtained as described using Beagle (Browning and Browning, 2007) version 3.3.2. The command line used for imputation is "java -Xmx10 g -jar beagle.jar unphased = sample.bgl missing = N out = imputed." The imputed genotype dataset was converted to BED format using PLINK (Purcell et al., 2007) version 1.07. We carried out GWAS with GEMMA (Zhou and Stephens, 2012), which uses a linear mixed model for association tests, using an estimating a relatedness matrix as a covariate. For real rice data, we further controlled population structure by using the first four principal components from PCA (see Population Structure Analysis of the Rice Mini-Core Collection) as covariates. Genome-wide critical values were determined by permutations: each studied phenotype was permuted for 200 times; for each permuted phenotype, GWAS was conducted and the genome-wide lowest P value was recorded. We then took the 5% lowest tail from the 200 recorded minimal P values as the threshold for genome-wide significance. The Manhattan and QQ plots for GWAS were generated using the R package qqman (Turner, 2014).

Rice Material and Sequencing

The Rice Mini-Core Collection was requested from USDA-ARS Genetic Stock Oryza Collection (http://www.ars.usda.gov/Main/Docs.htm?docid=23695, accessed March 7, 2016). All accessions had been purified as single seed decent (Li et al., 2010). Received seeds were compared with database photos for identification. They were then planted in Hainan (October 2011 to June 2012) for identification and phenotyping. In brief, germinated seeds were disseminated on seedbeds evenly so that they grew into the seedling stage under normal field management. After approximately 20 days, rice seedlings were transplanted to specifically designed rice plots with row distance $\geq$35 cm and plant distance $\geq$20 cm within a row. To avoid cross-pollination, we bagged each rice line at the top before flowering. The rice samples were further identified and confirmed at the flowering stage and harvesting. After confirmation, rice seeds from a single line were incubated in a plant growth chamber at 30°C for 10 h (day/light) and at 28°C for 14 h (night/dark) for 2 weeks. Young seedlings were harvested and used for further genomic DNA preparation. We adopted a CTAB method to extract DNA. In brief, ~2 g of fresh seedling material was soaked and grounded into fine powder in liquid nitrogen and thoroughly mixed with 2× CTAB DNA extraction buffer (100 mM Tris-HC1 [pH 8.0], 1.4 M NaCl, 20 mM EDTA [pH 8.0], 2% CTAB). After incubation at 28°C for 30 min, DNA was extracted with equal volume of chloroform and precipitated with 0.8 volume of isopropanol. Genomic DNA was washed twice with 99% ethanol before being dissolved in 100 μl of water. We determined the DNA concentration with Nanodrop 2000 (Thermo Scientific, Waltham, MA), and no less than 3 μg of DNA for each sample was used for sequencing library construction. DNA libraries (400–500 bp) were prepared and sequenced with a Hiseq2000 genome analyzer (Illumina, San Diego, CA) following manufacturer's instructions, with 90-bp paired end reads generated. The raw sequence data were further processed by removing adaptors and low-quality reads (more than 50% of bases have quality $\leq$5). The library preparation, genome sequencing, and raw data processing were conducted in BGI-Shenzhen, China.

Polymorphism Detection in Rice Population

Clean reads were mapped to the reference rice genome (Kawahara et al., 2013) (IRGSP-1.0) with BWA version 0.7.0 (Li and Durbin, 2009) using default parameters. With the mapped reads, we conducted remapping using Stampy version 1.0.20 (Lunter and Goodson, 2011). PCR duplicates were removed by "rmdup" in SAMtools version 0.17 (Li et al., 2009). To reduce miscalls caused by misalignment in INDEL regions, we realigned reads at the gapped region with GATK version 2.6 (DePristo et al., 2011). Before SNP calling, we masked all repeat regions in the rice genome to avoid false SNPs caused by mapping errors. As the false positives in SNP calling may further introduce errors in all downstream analysis, we applied stringent filtering. A repeat sequence database was created by combining the repeat database from RAP-DB (http://rapdb.dna.affrc.go.jp/) and that from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). In total, 179 Mb of the rice genome was masked from downstream analysis. From the non-repetitive region of the genome, we detected polymorphic sites with ANGSD using the previously described protocol. We further filtered the SNP dataset by removing all sites with a missing rate greater than 60%. These efforts led to a SNP dataset of 2 288 867 sites. Genotypes were called at these sites, and the accuracy of the genotype dataset was evaluated by Sanger sequencing. In brief, we used seeds from the same plants as used for genome sequencing for samples from each of the *japonica*, *indica*, and *aus* subgroups, to account for artifacts caused by mapping of divergent genomes to the same reference genome. DNA was extracted from young seedlings, fragmented into 300–12 000 bp, purified by gel extraction of 500–700 bp fragments, and inserted into TA plasmids (pMD 18-T; TaKaRa, Japan). The ligation products were transformed into homemade competent *Escherichia coli* DH5α. After culture and PCR screening, positive strains with inserted fragment size of 500–700 bp were sent to BGI-Shenzhen for plasmid isolation and sequencing (3730 DNA Analyzer; Applied Biosystems, USA). For each sample, we sequenced 100–200 random clones. The sequences were then cleaned by removing vector sequences and the T and A nucleotides at the 5′ and 3′ end, which were introduced by the TA cloning. The clean sequences were mapped to the rice genome to call genotypes in order to evaluate the accuracy of genotypes called from the Illumina data. The results are summarized in Supplemental Table 2.

Population Structure Analysis of the Rice Mini-Core Collection

We adopted genotype likelihoods-based methods for population structure analysis (Fumagalli et al., 2014). These methods avoid biases associated with genotype calling and are especially suitable for analysis of low-coverage genomic data. We reduced genome-wide SNP marker redundancy by randomly picking one SNP from every 5-kb region, which yielded a SNP dataset consisting of 52 838 sites evenly distributed across the rice genome. Genotype likelihoods at the these sites were estimated for downstream analysis using "-doGlf 2" and "-GL 1" arguments (Li, 2011) in ANGSD (version 0.542). PCA was conducted with ngsCovar from the ngsTools (Fumagalli et al., 2014) package. We excluded sites with MAF less than 5% by using "-minmaf 0.05," and we also disabled normalization as suggested in Patterson et al. (2006) by applying "-norm 0." Structure analysis was performed using NGSadmix (Skotte et al., 2013), which estimates individual ancestry directly from genotype likelihoods. We ran NGSadmix by varying K from 2 to 7. For each K, we ran NGSadmix for 200 replications with different seeds, and the clustering model with highest likelihoods was selected. Based on the clustering model with $K = 5$, we assigned each individual to one of the five subgroups of *Oryza sativa* using the following rule: accessions with $\geq$80% genetic ancestry derived from one of the five populations will be assigned to this subgroup, otherwise they will be assigned as *admixed*. To calculate the inbreeding coefficient for each individual, we used ngsF (Vieira et al., 2013) with default parameters and an input genotype likelihoods file produced by ANGSD using argument "-doGlf 0". All plots were created with R version 3.0.2.

Population Genetic Analysis at the Associated Loci

The local LD pattern at the *Wx* gene region was calculated by including 125-kb regions from both sides of the gene. In total, 1527 SNP markers were included, and imputed genotypes were obtained to construct the *Wx* haplotypes. The squared allele-frequency correlation (r^2) for all pairs of SNPs was determined with Haploview (Barrett et al., 2005). To reveal the population genetic pattern at the *qPc10* locus, we divided all *aus* samples into two subpopulations according to their allelic states at site Chr10:2091265. We estimated the diversity at the *qPc10* locus for AUS-A and AUS-G using the average number of pairwise differences (θ_π) calculated using "-doThetas" in ANGSD. The approach in ANGSD accommodates missing data and inherent uncertainty in low-coverage data and produces estimates comparable with those that would have been obtained from true genotypes without uncertainty (Korneliussen et al., 2014). To test the genome-wide significance of the diversity reduction for AUS-A, we retrieved all non-redundant polymorphic sites with MAF $\geq$0.4 (comparable with the MAF of the site Chr10:20912658, which is 50%) in the *aus* population. For each site, the *aus* population was divided into two subpopulations according to the ancestral/derived information and the ratio of θ_πs was calculated for these two subpopulations in a 20-kb region around Chr10:20912658. The ancestral/derived state was inferred by comparisons to the *Oryza punctata* genomic sequence (Wang et al., 2014). The genetic distance between the two populations, calculated as described by He et al. (2011), measures the average distance for all pairwise comparisons between sequences randomly drawn from the two populations, and ranges from 0 to 1. To evaluate the genome-wide significance of genetic distance between alleles at the *qPc10* region, we adopted the same procedure as described for the diversity reduction and calculated the genetic distance between the two subpopulations harboring different alleles. The distribution of the genetic distances is plotted in Supplemental Figure 14.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This research was supported by 973 Basic Research Project from the Ministry of Science and Technology of China (2015CB150106), the National Natural Science Foundation of China (31430063), and the State Key Laboratory of Plant Genomics.

AUTHOR CONTRIBUTIONS

C.C., R.N., and J.W. supervised the project. H.W. performed the simulation studies. X.X., H.W., and F.V. generated and processed the raw sequencing data, and performed the preliminary bioinformatics analysis to call genotype likelihoods and SNPs. H.W., Y.X., and Z.L. did the genome-wide association studies. H.W. performed the majority of the population genetic analysis with contributions from F.V. and R.N. H.W. and R.N. wrote the manuscript with critical input from all the authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Gupo Li from C.C.'s laboratory for the assistance in field management of the plant materials. No conflict of interest declared.

Received: January 7, 2016
Revised: April 16, 2016
Accepted: April 27, 2016
Published: May 11, 2016

REFERENCES

Agrama, H., Yan, W., Lee, F., Fjellstrom, R., Chen, M.H., Jia, M., and McClung, A. (2009). Genetic assessment of a mini-core subset developed from the USDA rice genebank. Crop Sci. **49**:1336–1346.

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

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

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. **81**:1084–1097.

Bryant, R., Proctor, A., Hawkridge, M., Jackson, A., Yeater, K., Counce, P., Yan, W., McClung, A., and Fjellstrom, R. (2011). Genetic variation and association mapping of silica concentration in rice hulls using a germplasm collection. Genetica **139**:1383–1398.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T.L., Polato, N.R., Olsen, K.M., Nielsen, R., McCouch, S.R., et al. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. **3**:1745–1756.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat. Genet. **46**:714–721.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**:491–498.

Ewing, G., and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**:2064–2065.

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

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS One **8**:e79667.

Fumagalli, M., Vieira, F.G., Linderoth, T., and Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics **30**:1486–1487.

Furukawa, T., Maekawa, M., Oki, T., Suda, I., Iida, S., Shimada, H., Takamure, I., and Kadowaki, K.i (2007). The *Rc* and *Rd* genes are involved in proanthocyanidin synthesis in rice pericarp. Plant J. **49**:91–102.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. USA **93**:10274–10279.

He, Z., Zhai, W., Wen, H., Tang, T., Wang, Y., Lu, X., Greenberg, A.J., Hudson, R.R., Wu, C.I., and Shi, S. (2011). Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet. **7**:e1002100.

Hindorff, L.A., Junkins, H.A., Hall, P., Mehta, J., and Manolio, T. (2016). A Catalog of Published Genome-wide Association Studies. http://www.genome.gov/gwastudies.

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

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

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012a). ART: a next-generation sequencing read simulator. Bioinformatics **28**:593–594.

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

Izawa, T., Konishi, S., Shomura, A., and Yano, M. (2009). DNA changes tell us about rice domestication. Curr. Opin. Plant Biol. **12**:185–192.

Jia, L., Yan, W., Zhu, C., Agrama, H.A., Jackson, A., Yeater, K., Li, X., Huang, B., Hu, B., McClung, A., et al. (2012). Allelic analysis of sheath blight resistance with association mapping in rice. PLoS One **7**:e32703.

Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., Chai, Y., Yang, L., Liu, K., Lu, H., et al. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nat. Genet. **45**:957–961.

Kato, S., and Ishikawa, J. (1921). On the inheritance of the pigment of red rice. Jpn. J. Genet. **1**:1–7.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice **6**:4.

Konishi, S., Ebana, K., and Izawa, T. (2008). Inference of the japonica rice domestication process from the distribution of six functional nucleotide polymorphisms of domestication-related genes in various landraces and modern cultivars. Plant Cell Physiol. **49**:1283–1293.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics **15**:356.

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

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079.

Li, X., Yan, W., Agrama, H., Hu, B., Jia, L., Jia, M., Jackson, A., Moldenhauer, K., McClung, A., and Wu, D. (2010). Genotypic and phenotypic characterization of genetic differentiation and diversity in the USDA rice mini-core collection. Genetica **138**:1221–1230.

Li, X., Yan, W., Agrama, H., Jia, L., Shen, X., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., and Wu, D. (2011a). Mapping QTLs for improving grain yield using the USDA rice mini-core collection. Planta **234**:347–361.

Li, Y., Sidore, C., Kang, H., Boehnke, M., and Abecasis, G. (2011b). Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. **21**:940–951.

Li, X., Yan, W., Agrama, H., Jia, L., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., and Wu, D. (2012). Unraveling the complex trait of harvest index with association mapping in rice (*Oryza sativa* L.). PLoS One **7**:e29350.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**:43–50.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. **21**:936–939.

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

Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera Lizarazu, O., Brown, P.J., Acharya, C.B., Mitchell, S.E., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc. Natl. Acad. Sci. USA **110**:453–458.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2011). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One **7**:e37558.

Olsen, K.M., Caicedo, A.L., Polato, N., McClung, A., McCouch, S., and Purugganan, M.D. (2006). Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. Genetics **173**:975–983.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics **26**:419–420.

Pasaniuc, B., Rohland, N., McLaren, P., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B., Daly, M., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. **44**:631–635.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. **2**:e190.

Pinson, S.R.M., Tarpley, L., Yan, W., Yeater, K., Lahner, B., Yakubova, E., Huang, X.-Y., Zhang, M., Guerinot, M.L., and Salt, D.E. (2015). Worldwide genetic diversity for mineral element concentrations in rice grain. Crop Sci. **55**:294–311.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81**:559–575.

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. **22**:939–946.

Shirley, B. (1998). Flavonoids in seeds and grains: physiological function, agronomic importance and the genetics of biosynthesis. Seed Sci. Res. **8**:415–422.

Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics **195**:693–702.

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

Sweeney, M.T., Thomson, M.J., Cho, Y.G., Park, Y.J., Williamson, S.H., Bustamante, C.D., and McCouch, S.R. (2007). Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet. **3**:e133.

The 3,000 Rice Genomes Project. (2014). The 3,000 rice genomes project. GigaScience **3**:7.

Tian, Z.X., Rizzon, C., Du, J.C., Zhu, L.C., Bennetzen, J.L., Jackson, S.A., Gaut, B.S., and Ma, J.X. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. **19**:2221–2230.

Turner, S.D. (2014). qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. bioRxiv, 005165. http://dx.doi.org/10.1101/005165.

Vieira, F.G., Fumagalli, M., Albrechtsen, A., and Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. Genome Res. **23**:1852–1861.

Wang, Z.Y., Zheng, F.Q., Shen, G.Z., Gao, J.P., Snustad, D.P., Li, M.G., Zhang, J.L., and Hong, M.M. (1995). The amylose content in rice endosperm is related to the post-transcriptional regulation of the *Waxy* gene. Plant J. **7**:613–622.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S., et al. (2014). The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. **46**:982–988.

Wei, X., Liu, K., Zhang, Y., Feng, Q., Wang, L., Zhao, Y., Li, D., Zhao, Q., Zhu, X., Zhu, X., et al. (2015). Genetic discovery for oil production and quality in sesame. Nat. Commun. **6**:8609.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., et al. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun. **5**:3438.

Yan, W.G., Li, Y., Agrama, H.A., Luo, D., Gao, F., Lu, X., and Ren, G. (2009). Association mapping of stigma and spikelet characteristics in rice (*Oryza sativa* L.). Mol. Breed **24**:277–292.

Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., Tyagi, W., Ali, M.L., Tung, C.W., Reynolds, A., et al. (2010). Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. PLoS One **5**:e10780.

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

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44**:821–824.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. **33**:408–414.