

Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm

Xuehui Huang^{1,2,5}, Yan Zhao^{1,2,5}, Xinghua Wei^{3,5}, Canyang Li¹, Ahong Wang¹, Qiang Zhao¹, Wenjun Li¹, Yunli Guo¹, Liuwei Deng¹, Chuanrang Zhu¹, Danlin Fan¹, Yiqi Lu¹, Qijun Weng¹, Kunyan Liu¹, Taoying Zhou¹, Yufeng Jing¹, Lizhen Si¹, Guojun Dong^{1,3}, Tao Huang¹, Tingting Lu¹, Qi Feng¹, Qian Qian³, Jiayang Li⁴ & Bin Han^{1,2}

A high-density haplotype map recently enabled a genome-wide association study (GWAS) in a population of *indica* subspecies of Chinese rice landraces. Here we extend this methodology to a larger and more diverse sample of 950 worldwide rice varieties, including the *Oryza sativa indica* and *Oryza sativa japonica* subspecies, to perform an additional GWAS. We identified a total of 32 new loci associated with flowering time and with ten grain-related traits, indicating that the larger sample increased the power to detect trait-associated variants using GWAS. To characterize various alleles and complex genetic variation, we developed an analytical framework for haplotype-based *de novo* assembly of the low-coverage sequencing data in rice. We identified candidate genes for 18 associated loci through detailed annotation. This study shows that the integrated approach of sequence-based GWAS and functional genome annotation has the potential to match complex traits to their causal polymorphisms in rice.

Rice (*Oryza sativa* L) is an economically important crop that accounts for ~20% of the world's caloric intake. To be grown successfully under a variety of climatic conditions across the globe, breeders maintain rice at high genetic diversity. Second-generation sequencing technologies have enabled resequencing of a large number of genomes and have provided the possibility of high-throughput genotyping and large-scale genetic variation surveys¹. Identification of allelic variations underpinning the phenotypic diversity observed in rice will have enormous practical implications in rice breeding².

Recently we performed low-coverage sequencing of 517 Chinese rice landraces and imputed missing genotypes to construct a haplotype map of the rice genome. We then used this map to perform GWAS in the *indica* population in a previous study³. However, identifying the loci associated with complex traits in rice is challenging. This is because (i) *O. sativa* contains *indica* and *japonica* subspecies, which can be further divided into several divergent groups with high amounts of population differentiation^{2,3}, and (ii) there is a low rate of linkage disequilibrium (LD) decay in rice³. Therefore, the interpretation of association signals and the identification of causal genes through GWAS requires a full incorporation of population structure and detailed follow-up analyses of associated loci for candidate genes and causal polymorphisms.

In this study, we examined 950 worldwide rice cultivars, representing a much broader and larger sample than has previously been used. We developed a new analytical framework to assemble low-coverage

sequences of different gene alleles⁴. This approach was then used to detect SNPs and complex polymorphisms such as insertions and deletions (indels). Using this new method, we were able to project a map of genic variation onto the rice genome. This facilitated the discovery of functional variation among rice varieties. We collected phenotypic data of flowering time (heading date) and grain-related traits and used them for a GWAS in the *O. sativa indica* and *japonica* subpopulations and in the full *O. sativa* population. The broader sampling greatly enhanced the power of the GWAS. In addition to the loci identified previously, we identified 32 new loci underlying flowering time and ten grain-related traits. In the follow-up analysis of these regions, we integrated detailed annotation, expression profiles and genetic variation to identify candidate genes and potential causal polymorphisms for the grain-related traits.

RESULTS

Genetic structure of worldwide rice germplasm

The germplasm collection used in this study included a previous set of 520 Chinese landraces, plus a new set from China and other widespread countries (**Supplementary Fig. 1**). We sequenced the genomes of the plants in the new set, which included 100 additional Chinese *japonica* landraces and 330 diverse global cultivars from 33 countries, on the Illumina Genome Analyzer IIx to approximately onefold coverage. The resulting sequence dataset of 950 rice varieties consisted of 4.6 billion 73-bp paired-end reads. After aligning these short reads

¹National Center for Gene Research, National Center for Plant Gene Research (Shanghai), Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. ²Chinese Academy of Sciences Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. ³State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China. ⁴National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. ⁵These authors contributed equally to this work. Correspondence should be addressed to B.H. (bhan@ncgr.ac.cn).

Received 3 June; accepted 2 November; published online 4 December 2011; doi:10.1038/ng.1018

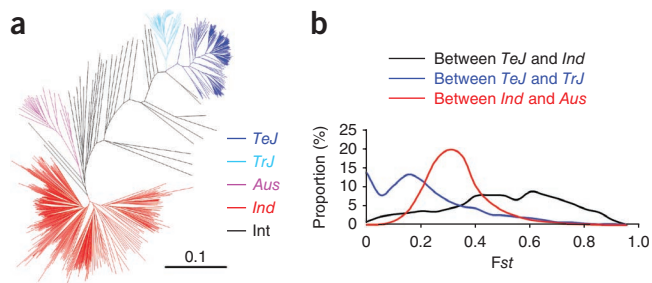


Figure 1 Genetic structure and population differentiation in 950 rice accessions. **(a)** Neighbor-joining tree of 950 rice accessions constructed from a simple matching distance of 4.1 million SNPs. The five divergent groups, *indica* (*Ind*), *aus* (*Aus*), *temperate japonica* (*TeJ*), *tropical japonica* (*TrJ*) and intermediate (*Int*), are colored in red, purple, blue, cyan and black, respectively. The scale bar indicates the simple matching distance. **(b)** The distributions of the pairwise population-differentiation statistic (*Fst*) across the rice genomes between *indica* and *temperate japonica* (in black), between *temperate japonica* and *tropical japonica* (in blue) and between *indica* and *aus* (in red).

to the *japonica* Nipponbare sequence (International Rice Genome Sequencing Project (IRGSP) 4.0) and SNP calling, we identified a total of 4,109,366 non-singleton SNPs. We calculated the allele frequency for each SNP site and found ~44% of the SNPs to be of low frequency (minor allele frequency (MAF) < 0.05 across the sample; **Supplementary Fig. 2**). The proportion of rare SNPs we examined, although similar to those reported in the human and *Arabidopsis* genomes^{5,6}, may have been underestimated because of the low power to call singleton variants from our low-coverage sequencing data.

We then investigated the population structure of the worldwide rice germplasm collection. We constructed a neighbor-joining tree and performed principal component analysis based on all the SNPs in the collection. The results showed that the collection has five divergent groups: *indica*, *aus*, *temperate japonica*, *tropical japonica* (also named *javanica*) and intermediate (a mix of more than one of the other four groups) (**Fig. 1a** and **Supplementary Fig. 3**). Among them, *indica* and *aus* cluster within the *indica* subspecies, whereas *temperate japonica* and *tropical japonica* are within the *japonica* subspecies. The *tropical japonica* and *aus* accessions are mainly grown in southeast Asia and south Asia, respectively (**Supplementary Table 1**), and we found that no traditional Chinese landraces belong to these two groups. Compared to the previous collection of 520 Chinese landraces, this worldwide collection contains many diverse *indica* and *temperate japonica* varieties that are found in a wide range of geographical, climatic and cultural conditions (**Supplementary Figs. 4** and **5**).

Based on the SNP data, we estimated the sequence diversities (π) of the three subpopulations *tropical japonica*, *temperate japonica* and *aus* to all be 0.0006, which is much lower than the sequence diversity of *indica* ($\pi = 0.0016$). According to pairwise estimates of the population-differentiation statistic (*Fst*), the average amount of population differentiation between *tropical japonica* and *temperate japonica* was lower than that between *indica* and *aus*, whereas the differentiation between the *indica* and *japonica* subspecies was much greater (**Fig. 1b**). The sequence diversity (π) of each group, calculated in 100-kb windows across the rice genome, and the pairwise population differentiation (*Fst*) between any two groups are shown in **Supplementary Tables 2** and **3**, respectively. We scanned the groups for highly differentiated loci and found seven loci between *tropical japonica* and *temperate japonica* and six loci between *indica* and *aus* with *Fst* > 0.8, indicating that these loci might be involved in adaptation or might be a result of founder effects (**Supplementary Table 4**).

We calculated the allele frequency of each SNP site in four populations (*indica*, *aus*, *temperate japonica* and *tropical japonica*) and found that >10% of the SNPs had large allele frequency differences in the pairwise population comparison. Among these SNPs, we found a total of 315 sites with population-specific alleles (that is, with an allele frequency >0.95 in one population and <0.05 in the other three), of which there were seven nonsynonymous population-specific SNPs, all located within the highly differentiated loci identified above. Detailed annotations of the highly differentiated loci and population-specific SNPs are provided in the **Supplementary Note** and **Supplementary Table 5**.

We used the *k*-nearest neighbor algorithm for data imputation of the genotype dataset of the 950 varieties. We also performed the imputation specifically on the subset of 508 *indica* accessions (given the population differentiation, we did not put *aus* accessions into the subset for the imputation) and the subset of 383 *japonica* accessions (including both *tropical japonica* and *temperate japonica*) to perform GWAS for each subspecies. The large number of accessions sequenced improved the imputation efficiencies. The accuracy of genotype calls was above 99% after imputation (**Supplementary Note** and **Supplementary Table 6**). The resulting genotype datasets contained a total of 1,345,417 SNPs (MAF > 0.03) in the full population set, 743,014 SNPs (MAF > 0.05) in the subset of *indica* and 493,777 SNPs (MAF > 0.05) in the subset of *japonica*.

Local *de novo* sequence assembly for causal variant detection

Low-coverage second-generation sequencing coupled with an effective imputation algorithm can yield a high-resolution genotype dataset for GWAS in rice. However, a remaining issue is whether low-fold sequence coverage on a large number of samples can also be used to detect more complex polymorphisms such as indels⁴. Unlike array-based genotyping methods, second-generation resequencing has the potential to detect extensive new variations directly⁷. To overcome drawbacks resulting from short read lengths and low sequencing coverage, we developed a haplotype-based local sequence assembly method, which combines the sequencing reads of a common haplotype in a local region to perform *de novo* assembly.

The analytical framework of the local sequence assembly is as follows: (i) the samples for GWAS are divided into several common haplotype groups based on the peak SNPs or the local genotype in the associated loci; (ii) the sequence read pairs around the associated loci of each group are then combined for *de novo* assembly; and (iii) given the high SNP density from sequencing and the low rate of LD decay in the rice genome, nearly complete LD between peak SNPs and causal polymorphisms should be found. Hence, against the background of other segregating sequences, the causal polymorphisms can potentially be found in the consensus sequences of the haplotype groups by comparison of the assembled sequences with the reference genome sequences. This approach facilitates the detection of potential causal variants in follow-up analyses of GWAS, which have typically included PCR amplification of target regions followed by resequencing.

We applied this approach to detect causal polymorphisms in six known loci identified previously (loci in *Waxy*, *ALK*, *Rc*, *OsC1*, *GS3* and *qSW5*)³. Local assembly generated one to two contigs for each allele, covering nearly the entire target regions of each (**Fig. 2**). We then analyzed the assembled sequences and searched for functional variations. We found a 23-bp duplication in *Waxy* (resulting in a premature stop codon; **Fig. 2a**), a 2-bp difference in *ALK* (resulting in a nonsynonymous mutation from phenylalanine to leucine; **Fig. 2b**), a 14-bp deletion in *Rc* (resulting in a frameshift mutation; **Fig. 2c**), a 10-bp deletion in *OsC1* (resulting in a frameshift mutation; **Fig. 2d**) and a 1-bp

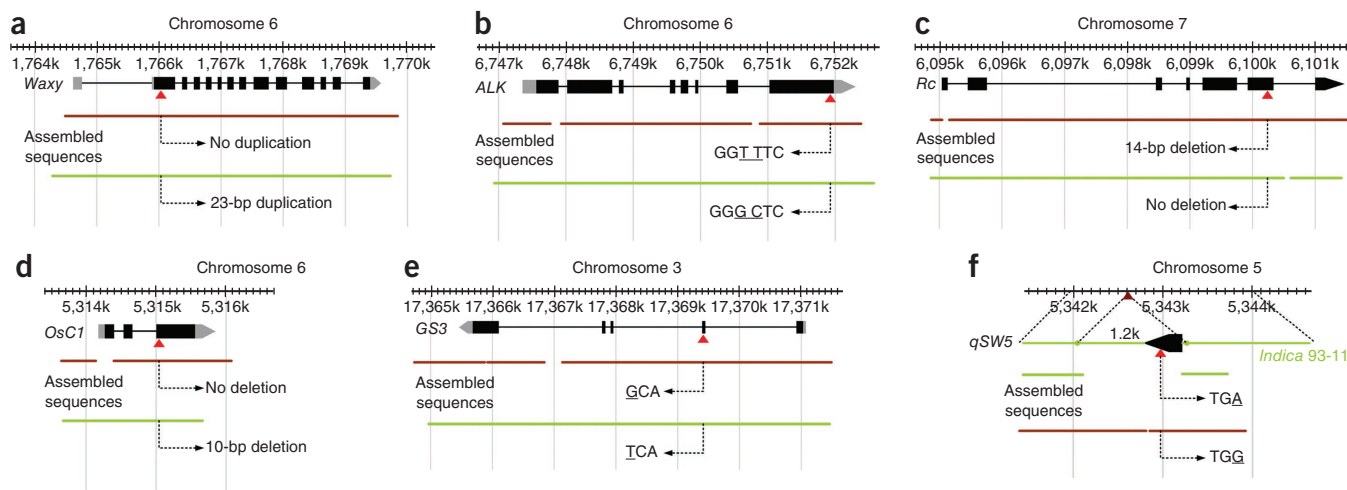


Figure 2 Causal variant detection in six genes previously identified by GWAS in the *indica* population. The top of each panel shows the genomic location of the known gene and its gene structure. Exons and introns are depicted as rectangles and lines, respectively. Coding regions and untranslated regions are shown in black and gray, respectively. The points indicate the orientation of the genes. The red triangles indicate the location of causal variants detected in the gene. The bottom of each panel shows the contigs of two alleles (with major allele shown as a brown line and the minor allele shown as a green line) from local assembly, where the genotypes in the causal variant site are indicated. (a) *Waxy* is responsible for amylose content. (b) *ALK* is responsible for starch gelatinization temperature. (c) *Rc* is responsible for pericarp color. (d) *OsC1* is responsible for apiculus color. (e) *GS3* is responsible for grain length. (f) *qSW5* is responsible for grain width. The dark red triangle indicates the location of 1.2-kb deletion in the reference genome. The *indica* cv. 93-11, which has sequences from whole-genome shotgun and belongs to the haplotype group with the minor allele in *qSW5*, was used as a template to compare to assembly results.

substitution in *GS3* (resulting in a premature stop codon; **Fig. 2e**). The functional variants detected here are almost all consistent with previous findings^{8–12}. However, *qSW5* (also called *GW5*) was an exception. This is because the *japonica* Nipponbare reference genome harbors a large deletion in this gene^{13,14}. The local assembly results showed that both haplotype groups of *qSW5* in *indica* do not contain such a deletion (**Fig. 2f**). The only difference in the coding region of *qSW5* between the two haplotype groups in *indica* was a nonsynonymous mutation (from proline to serine in the nuclear localization site of the protein), which is probably the causal polymorphism of *qSW5* in the *indica* population. Together, these results show that the haplotype-based sequence assembly method is a powerful approach for detecting putative causal polymorphisms.

Common complex polymorphisms in rice genes

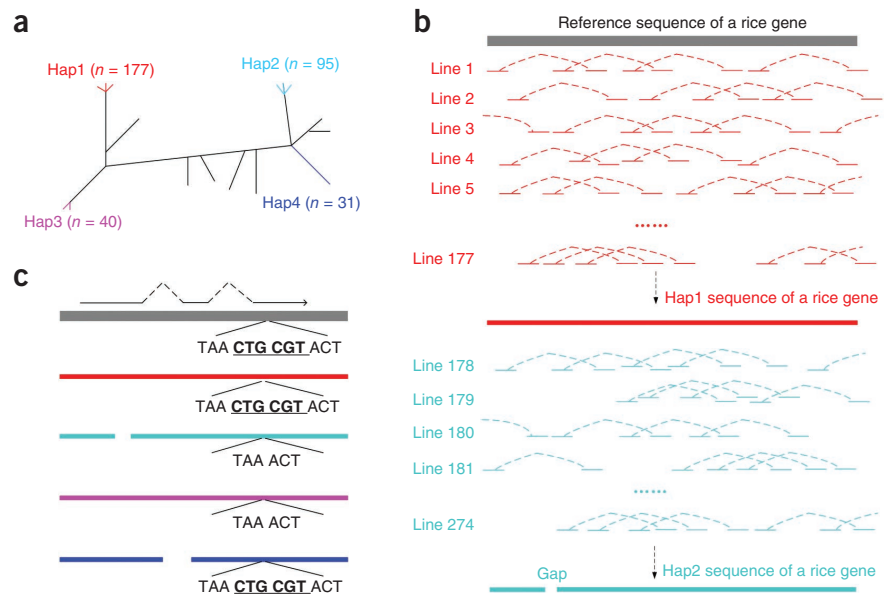
We extended these methods to identify common complex polymorphisms in all rice genes that have evidence of transcription. We grouped the Chinese rice landraces (both *indica* and *japonica*) into several haplotypes according to the local genotypes around each gene (**Fig. 3** and **Supplementary Note**). We found that ~92% of rice genes have at least five haplotype groups, with a mean of 3.6 haplotype groups per gene (**Supplementary Fig. 6a**). The resulting assembly sequences have an average of 1.8 contigs for each gene within a haplotype group (**Table 1** and **Supplementary Fig. 6b**). We compared the assembly sequences against their corresponding bacteria artificial chromosome (BAC)-based sequences (*indica* cv. Guangluai-4 and *japonica* cv. Nipponbare) and found that the consistency of one member within the haplotype group between the consensus sequences and the exact sequences was ~99.83% (that is, an average of 1.7 bp discrepancies per kb). Considering that variants in genic regions are more likely to influence protein function than non-genic variants, we did not extend this approach to the entire genome in this study.

We then aligned the assembled sequences to the rice reference genome for variation detection. In total, we identified 191,476

non-redundant indels in genic regions, ranging from 1 bp to 376 bp in size. The small insert size of our Illumina sequencing library (~400 bp) limited the power to assemble entire regions for direct identification of larger indels, but the partial sequences at the end of the insertions from the assembly could potentially be useful for the detection of larger variants (**Supplementary Fig. 7**). We also identified 660,909 non-redundant SNPs, of which 36.9% were new when comparing them with ~4.1 million SNPs above. The increase in sensitivity using this method was mainly a result of its power to call SNPs around highly polymorphic regions (for example, SNPs adjacent to indels). Based on the BAC-based sequences of two rice accessions (*indica* cv. Guangluai-4 and *japonica* cv. Nipponbare), we estimated the specificity of the variant calling to be ~90%. This value is likely an underestimate because both breakpoint location and insertion sequences were required to achieve an exact match in our specificity calculations. The assembled sequences were used solely to detect functional variation, and the indels and extra SNPs identified were not used in the GWAS. Taken together, the assembly approach for population-scale low-coverage sequencing can partially meet the demand for individual-scale deep sequencing: the assembled sequence is the consensus sequence of the local haplotype group and enables the detection of common variants, but it cannot fully reflect the actual sequence of a specific line, which would require high-coverage sequencing followed by whole-genome *de novo* assembly¹⁵.

We further predicted the effects of the variants on coding genes. The frameshift-to-non-frameshift ratio for indels was 0.37 and the nonsynonymous-to-synonymous ratio for SNPs was 1.1 (**Supplementary Table 7**). The low frameshift-to-non-frameshift ratio was similar to those found in human and *Arabidopsis* genes, where indels in coding regions occur preferentially in multiples of 3 bp to avoid frameshifts^{15,16}. In our dataset, we observed causal polymorphisms in many genes (for example, small indels in *Phr1* (ref. 17) and multiple amino-acid changes in *SKC1* (ref. 18)). We then focused on the variants predicted to cause large effects, which we refer to as

Figure 3 Illustration of haplotype-based local assembly. (a) For each gene, genotypes for 55 SNP sites around the gene were retrieved from the genotype dataset. We calculated the simple matching distances for the 55 SNP sites and performed hierarchical cluster analysis by using the single linkage algorithm, which generated several haplotype groups (Hap1 through Hap4). The haplotype groups with a frequency of < 0.02 were excluded. (b) For each haplotype group, paired-end reads (shown as thick lines, with dashed lines connecting the read pairs) that were uniquely mapped onto the local region were collected together. In the case that only one of the paired-end reads was aligned, we picked up both of them. The sequence assembly was performed for each haplotype group separately, generating one (no gaps) or several contigs. (c) The contigs of different haplotype groups were aligned with the reference genome for sequence variation detection. The sequence variants detected, including SNPs and indels, were then used to predict the potential effects on the gene.



large-effect variants. We found a total of 5,692 large-effect variants in 3,537 annotated genes. A detailed list of all the large-effect variants is given in **Supplementary Table 8**. Among these variants, we found that frameshift mutations caused by indels occupied the largest proportion (at 72.6% of all large-effect variants) (**Supplementary Table 7**), which was similar to the findings in the human genome¹⁹.

We used information from gene annotation, orthologs of known genes in *Arabidopsis* and public expression profiles from microarray to identify the genes involved in flowering time, disease resistance, drought tolerance and salt tolerance (Online Methods). We found 8, 128, 138 and 90 genes with mutations having large effects to be potentially involved in flowering time, disease resistance, drought tolerance and salt tolerance, respectively (**Supplementary Fig. 8**), including many known genes (for example, *Hd1*, *Xa26*, *Pi2* and *Pi9*)^{20–22}. Thus, the assembled sequences and the extensive variation dataset obtained through large-scale genome sequencing should enable rapid forward genetics studies, as well as support studies using a reverse genetics approach.

GWAS of 11 traits in *indica*, *japonica* and the full population

We phenotyped the population of 950 lines for 11 agronomic traits (heading date and ten grain-related traits; see **Supplementary Note, Supplementary Table 9** and **Supplementary Fig. 9** for more details). Heading date is a complex trait that controls adaptation of rice cultivars to their local environment and is strongly affected by population structure. Grain-related traits are crucial for rice improvement in terms of yield and end use. The ten grain-related traits we evaluated here included three traits of grain quality (grain protein content, amylose content and starch gelatinization temperature), four of yield components (grain number, grain length, grain width and grain weight) and three of coloration (seed hull color, pericarp color and apiculus color). We used both a simple model and a compressed mixed linear model (MLM)²³ to identify association signals, which we examined in the *indica* population (a total of 508 lines), the *japonica* population (a total of 383 lines) and the full rice population (a total of 950 lines). We optimized all the parameters in the compressed MLM for each trait in the three GWAS panels (**Supplementary Note** and **Supplementary Figs. 10–14**). The Manhattan plots and quantile-quantile plots of the 11 traits are shown with the results from both the simple model and the compressed MLM in **Supplementary Figures 15–25**.

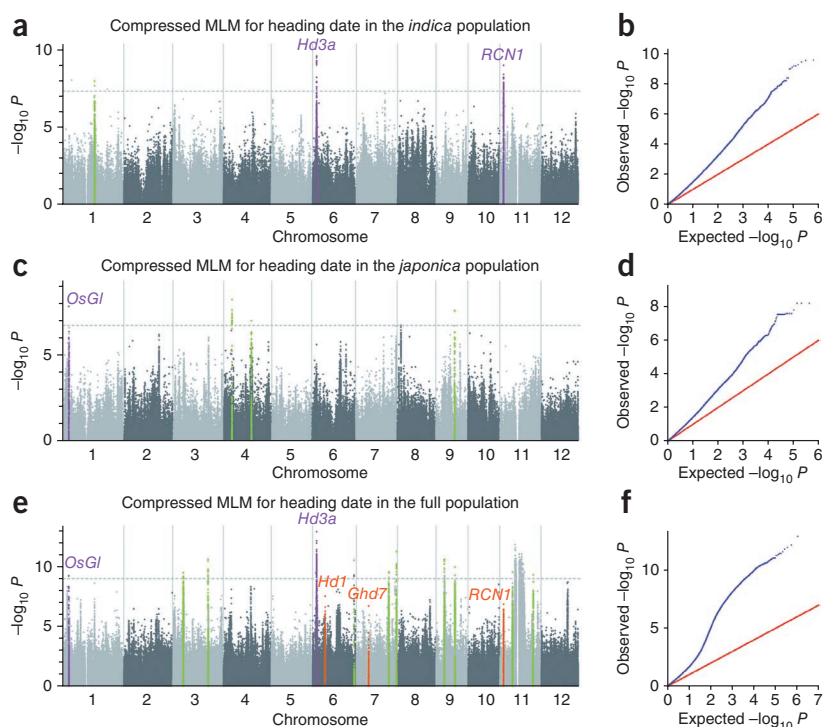
Compared to previous studies³, the power of our GWAS to detect associated loci was increased, likely as a result of substantial improvements in sample size and phenotyping. This effect can be partially observed in the GWAS of heading date—an extreme example in which the trait is strongly affected by population structure^{24,25}. Previous GWAS of 373 *indica* lines did not yield any known loci associated with heading date. Using our extended population, we conducted a GWAS on heading date in the three panels and detected significant association signals around known loci^{26,27} (**Fig. 4**). This result showed that our GWAS had the potential to dissect the genetic basis of complex traits such as heading date. Peak SNPs at the identified loci explained ~42% of the phenotypic variance in the *indica* population, ~46% of the variance in the *japonica* population and ~36% of the variance in the full population. In the outcrossing species maize, numerous small-effect quantitative trait loci have been identified for flowering-time variation²⁰. The genetic architecture of flowering-time variation in rice, which is a self-fertilizing plant species, is quite distinct, as heading-date variation appears to be controlled by a few genes of large effect.

Table 1 Summary of the haplotype-based local assembly in genic regions

Item	Statistic
Genes for haplotype-based assembly	
Rice genes with transcript support	27,784
Average length of rice genes (bp)	3,254
Average haplotypes per gene (MAF > 0.03)	3.6
Raw data for haplotype-based assembly	
Raw data used in assembly (Gb)	57.2
Average coverage (x)	158
Contigs from haplotype-based assembly	
Total length (Mb)	328.1
Total number	179,300
N50 length (bp)	3,377
Average length (bp)	1,830
Estimated error rate ^a	0.17%
Sequence variations detected from contigs	
Non-redundant SNPs in genic regions	660,909
Non-redundant indels in genic regions	191,476

^aThe rate was estimated based on the BAC sequences of *japonica* cv. Nipponbare and *indica* cv. Guangluai-4.

Figure 4 Genome-wide association study of heading date in the *indica* population, the *japonica* population and the full population using the compressed MLM. For the significant loci identified, known loci are shown in purple and newly discovered loci are shown in green. Of these loci, those in *Hd3a*, *Hd1* and *Ghd7* had their causal variants detected through the haplotype-based assembly method, and there were no variants detected in the coding regions of *RCN1* and *OsGI* (**Supplementary Note**). (a) Manhattan plots for heading date in *indica* population. The $-\log_{10} P$ values from a genome-wide scan are plotted against the position on each of the 12 chromosomes. The horizontal dashed line indicates the genome-wide significance threshold ($P = 5 \times 10^{-8}$). (b) Quantile-quantile plot for heading date in the *indica* population. The horizontal axis shows $-\log_{10}$ transformed expected P values, and the vertical axis indicates $-\log_{10}$ transformed observed P values. (c) Manhattan plots for heading date in the *japonica* population and the genome-wide significance threshold ($P = 2 \times 10^{-7}$, shown as a dashed line). (d) Quantile-quantile plot for heading date in *japonica* population. (e) Manhattan plots for heading date in the full population and the genome-wide significance threshold ($P = 1 \times 10^{-9}$, shown as a dashed line). (f) Quantile-quantile plot for heading date in the full population. Clear association signals around known genes that did not meet the genome-wide significance threshold are shown in orange.



The association signals of the grain-related traits were much clearer in the GWAS of 508 *indica* lines, which had less subpopulation differentiation (**Supplementary Fig. 4**). In addition to 12 associated loci identified previously using our smaller association panel, we also identified eight new loci in the *indica* population, most of which are associated with heading date and grain weight. A previous GWAS of grain weight did not identify any significantly associated loci³. We identified three new loci associated with grain weight. The second strongest signal was at *qSW5*, which has previously been reported to control both grain width and grain weight^{13,14}.

The power of GWAS in the *japonica* panel was relatively limited by its population size and population structure (**Supplementary Fig. 5**). We compared the phenotype data between the two *japonica* subspecies and found that there was evidence of strong differentiation at most of the traits (**Supplementary Fig. 26**). Such differences were also reflected in the genetic architecture of the traits, both in the associated loci detected and in their effect sizes. Notably, the genetic architecture of the coloration traits was quite similar between the two subspecies (**Supplementary Figs. 23–25**), the architecture of the grain yield traits was mostly similar (**Supplementary Figs. 20–22**) and the architecture of the grain quality traits was quite different (**Supplementary Figs. 15–17**). For example, we identified three associated loci underlying grain quality

in *japonica* panel, all of which were different from those in *indica* panel, whereas the three major loci for coloration were the same between the subspecies. The similarities in the coloration traits may be because visible traits are more likely to be affected by introgression between subspecies.

Table 2 A subset of associated loci and candidate genes according to gene annotation and expression profiles

Trait	Chr.	Position ^a	Candidate genes	Description			
Gelatinization temperature	7	6,404,473	<i>Os07g0213700</i> ^b ; <i>Os07g0214300</i> ^b ;	Gene cluster of multiple amylase inhibitors			
			<i>Os07g0215500</i> ^b				
Hull color	9	7,432,434	<i>Os09g0292900</i> ^b	<i>OsFBX310</i>			
Grain number	6	1,966,613	<i>Os06g0134700</i> ^b ; <i>Os06g0137100</i> ^b	Protein kinase-like; kinesin motor			
Grain width	2	35,545,283	<i>Os02g0812000</i> ; <i>Os02g0809900</i> ^b	Homolog of <i>AtCCR2</i> ; expressed protein			
Amylose content	10	19,100,315	<i>Os10g0490100</i> ; <i>Os10g0492600</i> ^b	<i>RIPER6</i> ; aquaporin protein			
			12		27,275,219	<i>Os12g0630200</i> ^b ; <i>Os12g0630500</i> ^b ;	
						<i>Os12g0630100</i> ^b	
Pericarp color	6	911,244	<i>Os06g0115000</i> ^b	Expressed protein			
			8		12,483,220	<i>Os08g0301500</i>	Sucrose-phosphate synthase
Grain protein content	2	27,066,892	<i>Os02g0650900</i>	Glutamate dehydrogenase			
			7		23,614,414	<i>Os07g0570100</i> ; <i>Os07g0570300</i> ;	
Grain weight	7	23,614,414	<i>Os07g0570500</i>	Gene cluster of three peptidase proteins			
			6		24,746,851	<i>Os06g0602800</i>	Glycosyl transferase
			11		4,343,017	<i>Os11g0184800</i>	<i>OsAsp1</i>
Grain length	7	19,248,564	<i>Os07g0497500</i> ^b	MYB transcription factor			
			2		5,049,913	<i>Os02g0192300</i> ^b	Zinc finger protein
Grain length	3	23,344,870	<i>Os03g0604566</i> ^b ; <i>Os03g0604600</i> ^b	Expressed protein; a transport protein			
			3		21,838,912	<i>Os03g0574600</i> ^b ; <i>Os03g0577100</i> ^b	Expressed protein; Nuf2 family protein
			6		9,282,785	<i>Os06g0266800</i> ^b ; <i>Os06g0265400</i> ^b	<i>GASR7</i> ; a transcription factor
	7	19,745,179	<i>Os07g0501800</i> ^b	Receptor-like kinase			

^aPosition in bp according to IRGSP 4.0. ^bThese genes had high expression in the specific stages. Chr., chromosome.

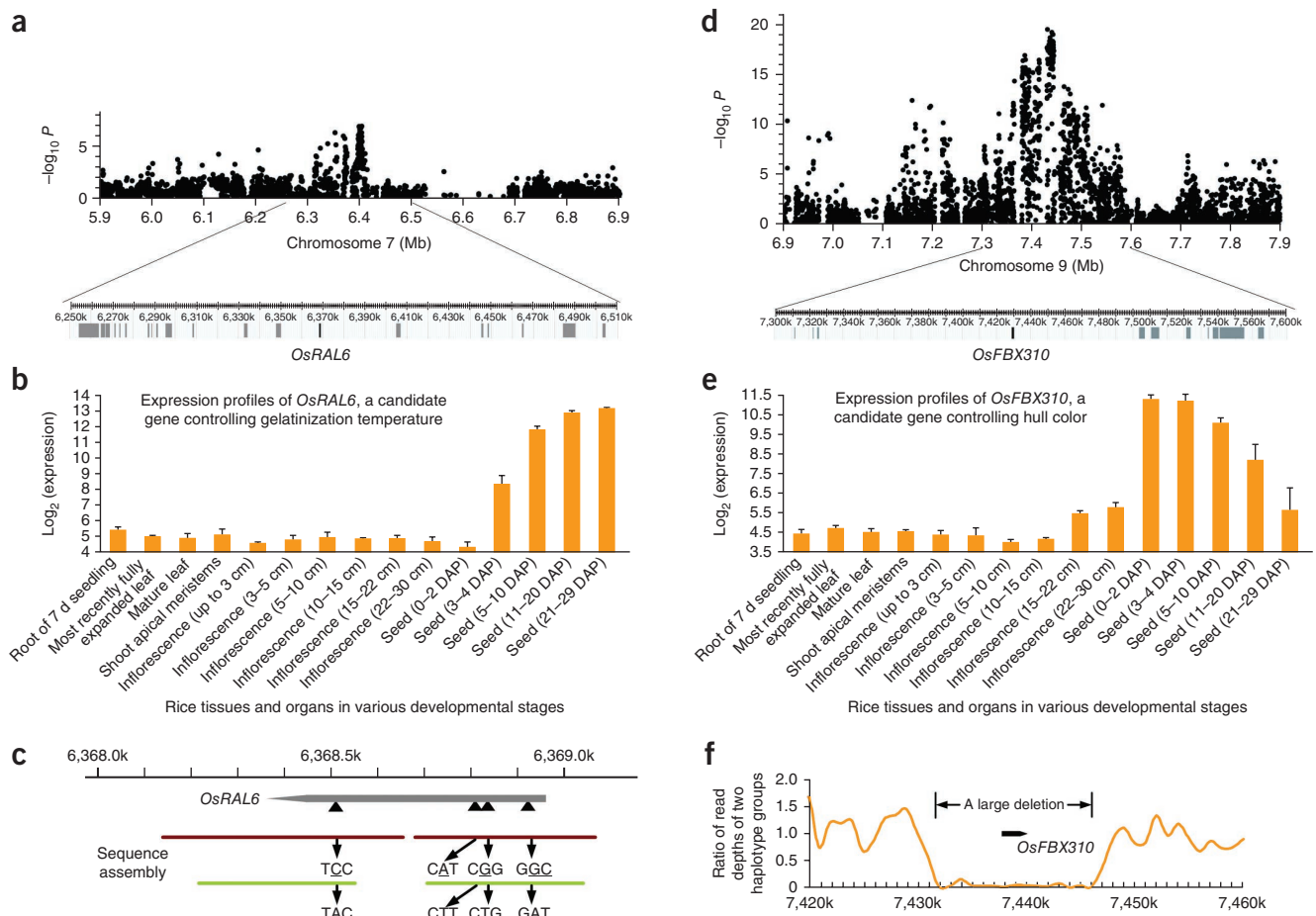


Figure 5 Regions of the genome showing association signals and the expression profiles of candidate genes. **(a)** An associated locus for starch gelatinization temperature. The top of the panel shows the region on each side of the peak SNP. $-\log_{10}$ transformed P values from the compressed MLM are plotted on the vertical axis. The bottom of the panel shows a narrow region, with the candidate genes indicated by dark gray. **(b)** The expression pattern of a candidate gene (*OsRAL6*) in the amylase inhibitor gene cluster from public microarray data. DAP, days after pollination. Error bars, s.d. of three replicates. **(c)** Functional variants detected through *de novo* local assembly. The triangles indicate the location of causal variants detected in the gene. The colored lines show the contigs of two alleles (with the major allele shown as a brown line and the minor allele shown as a green line) from local assembly, where the genotypes of four non-synonymous SNPs are indicated. **(d)** An associated locus for hull color. **(e)** Expression pattern of the candidate gene (*OsFBX310*). **(f)** The ratio of read depths (the ratio of the number of mapped reads from 41 *indica* rice lines with the minor allele to those from 41 *indica* lines with the major allele) are plotted against the local genomic region, and the positions of the candidate gene and the potential deletion are indicated.

A GWAS in the full population has higher power because of the larger population size, increased genetic diversity and increased SNP density. Most association signals detected from the GWAS in each subspecies could also be examined in the full population. However, a GWAS in the full population cannot fully replace those in each subspecies. Examining associations across two subspecies with strong population differentiation is feasible (Fig. 4e), although it is a challenge in statistical genetics (and probably produces a high number of false positives, as shown in Fig. 4f). In addition, some genes contain multiple different causal polymorphisms in the two subspecies (for example, *qSW5*; Fig. 2f), which reduces the power of the GWAS in the full population. We performed an association analysis for three traits (pericarp color, grain width and heading date) using haplotype information, which was not very successful at the whole-genome scale as compared to the association analysis performed using SNPs (Supplementary Figs. 27–29). These results indicate that further improvements to the statistical methodologies are needed in future studies.

To adopt appropriate thresholds to reduce the false-positive rate but also retain major associations, we tested different P value thresholds and chose thresholds based on quantile-quantile plots, known loci

and genome-wide type I error rates estimated through permutation tests (Supplementary Note). We identified association signals at $P < 5 \times 10^{-8}$ in the compressed MLM for *indica*, $P < 2 \times 10^{-7}$ for *japonica* and $P < 1 \times 10^{-9}$ for the full population. We retained as significant only those associations with clear peak-like signals in both the compressed MLM and the simple model. In total, our GWAS of the 11 traits identified 44 associated loci, 32 of which were new. Detailed information about the associated loci is shown in Supplementary Tables 10–12.

Follow-up analyses of associated loci for grain-related traits

We then selected candidate genes in the associated loci based on: (i) known involvement as metabolic or signaling genes in the corresponding traits and (ii) their expression profiles in rice. In fact, several known genes underlying these traits have distinct tissue-specific expression patterns. For example, we searched public expression profiles and literature descriptions for *Bh4* (controlling seed hull color)²⁸, *DEP1* (controlling grain number)²⁹, *GW2* (controlling grain weight and grain width)³⁰, *qSW5* (controlling grain width and grain weight)^{13,14}, *GS3* (controlling grain length)⁹, *Waxy* (controlling

amylose content)¹² and *ALK* (controlling starch gelatinization temperature)¹² and found that they all have a relatively high expression at specific stages of grain filling (or panicle development, for example, *DEP1*) and very low expression in the majority of the other temporal-spatial conditions (**Supplementary Note** and **Supplementary Fig. 30**). Hence, we used both expression profile data and gene annotation information to identify the putative function of genes around associated loci (**Supplementary Table 13**). We identified a total of 30 candidate genes in the peak SNP sites (or adjacent to these sites) of 18 associated loci (**Table 2**). The combined approach was not effective for all loci because of the complexity of candidate gene identification. In particular, the microarray expression data covered most but not all rice genes. In addition, publicly available expression profiles for a few rice cultivars cannot fully reflect the same conditions in GWAS for widely diverse varieties.

There are several association signals located in genomic regions with many tandemly repeated genes (**Table 2**). For example, in a newly identified locus controlling starch gelatinization temperature in the *japonica* population, there is a cluster of genes that can function as amylase inhibitors that all have very high expression in seeds at the ripening stages (**Fig. 5a,b**). *OsRAL6*, the gene immediately adjacent to the peak SNP for this trait, contains four amino acid changes between the associated haplotypes (**Fig. 5c**). Nevertheless, among multiple variants detected in this gene cluster, we could not identify the causal gene and the causal polymorphisms unambiguously because of the high LD in rice.

We also detected a large-effect variant in an associated locus for hull color (a locus also identified in a previous study³¹; **Fig. 5d**), which is probably the causal polymorphism. After screening the expression profiles of genes around the associated locus, only the gene immediately adjacent to the peak SNP, *OsFBX310* (*Os09g0292900*, encoding an F-box protein), showed extremely high expression enrichment in seeds at the ripening stages (**Fig. 5e**). We then assembled the two haplotype groups based on the peak SNP and found that the haplotype group with the minor allele (showing a brown hull) had very few unique reads in the region, probably as a result of a large deletion (**Fig. 5f**).

We reanalyzed raw sequence reads in the local region, trying to directly obtain genotypes at the causal polymorphism sites. We selected three genes (*Hd3a*, *OsRAL6* and *OsFBX310*) that all harbor predicted loss-of-function or non-synonymous mutations according to the haplotype-based assembly (**Supplementary Note**). We retrieved the genotype information for each rice variety according to its raw reads covering the putative causal polymorphism sites (that is, if the variety had those reads sequenced). We then examined whether predicted mutations do or do not relate to the phenotypic effects. The results (**Supplementary Tables 14–16**) revealed that the putative causal polymorphisms detected were in nearly complete LD with peak SNPs ($r > 0.8$). The polymorphisms all associated significantly with phenotypic variation but sometimes this association was slightly weaker than that of the peak SNPs (as has been previously observed³).

Hence, this study lays the foundation for studying the potential function of these candidate genes. Future work will use functional genomics methodologies such as genetic transformation and transferred DNA insertion mutant screens to validate the effect of these genes and their functional variants.

DISCUSSION

GWAS based on large-scale resequencing provide a powerful platform for finding genetic variants that can be directly used for crop

improvement^{32,33}. We reexamined loci identified by a previous smaller panel using 371 *indica* landraces³. For most traits, nearly all loci identified previously were still seen when using the larger panel. However, a few loci underlying heading date and grain number were identified by the smaller panel but were not found using the larger panel. This may be a result of phenotyping done in different years (hence, the environmental conditions may have been different) and statistic modeling for different population structures. We noted that most of the loci identified in the smaller panel that were not found in the larger panel are minor loci with *P* values just passing the significance thresholds, which are more likely to be affected by the factors above³⁴.

For future studies aiming to examine associations with particular traits, a case-control design may be required to maximize the probability of finding new associations. Some alleles are not under selection in common growth conditions and, therefore, their frequency is low in the population³⁵.

The low rate of LD decay in rice means that the resolution of GWAS cannot resolve a single gene, whereas ‘deep annotation’ of the genes in the local region can provide evidences for gene function. The gene expression patterns derived from high-throughput technologies (for example, expression microarray and RNA-Seq³⁶) provide important clues into the biological functions of genes. The expression results used in this study are all from previous independent experiments. The expression profiles of the diverse varieties of the collection in a particular tissue (or in a biotic or abiotic stress condition), if available, can more accurately aid in the determination of candidate genes. Such resources can also enable the identification of expression quantitative trait loci at high resolution, which would further our knowledge of the gene regulation network of the corresponding traits.

This study showed that local assembly enabled the identification of complex polymorphisms and facilitated the detection of causal variants in the associated loci^{4,37,38}. We found the performance of the method to decline slowly when genetic diversity increased and haplotype frequency decreased (**Supplementary Fig. 31**), and therefore, it should work well for organisms with relatively higher diversity and relatively more haplotypes. Individual-scale deep sequencing and population-scale low-coverage sequencing are two approaches that meet different needs and different goals¹. To reflect the actual sequences of specific lines, we are currently deep sequencing several diverse rice genomes and improving whole-genome *de novo* assembly algorithms, which aims to compensate for the limitation of the haplotype assembly method. In the future, more reference genome sequences in rice will greatly aid in the detection of causal variants.

URLs. All datasets of the worldwide collection of rice germplasm, <http://www.ncgr.ac.cn/RiceHap2/>; genotype dataset at the Rice Haplotype Map Project database, <http://www.ncgr.ac.cn/RiceHap2/Geno.html>; details of the 4.1 million SNPs, including their position and allele frequencies in each population, <http://www.ncgr.ac.cn/RiceHap2/SNPs.html>; local assembly sequences, complex variant data set (including detailed annotations) and BLAST searching, <http://www.ncgr.ac.cn/RiceHap2/Assembly.html>; the software used in this study, <http://www.ncgr.ac.cn/fimg/down.html>; IRGSP Build 4.0, <http://rgp.dna.affrc.go.jp/IRGSP/Build4/build4.html>; Rice Annotation Project Database (RAP-DB), <http://rapdb.dna.affrc.go.jp/>; Michigan State University Rice Genome Annotation Project (MSU-RGAP), <http://rice.plantbiology.msu.edu/>; Gene Expression Omnibus, <http://www.ncbi.nlm.nih.gov/geo/>; Ssaha2, <http://www.sanger.ac.uk/resources/software/ssaha2/>; Smalt, <http://www.sanger.ac.uk/resources/software/smalt/>; Fuzzypath, <ftp://ftp.sanger.ac.uk/pub/zn1/fuzzypath/>; MUMmer, <http://mummer.sourceforge.net/>;

EMBOSS, <http://emboss.sourceforge.net/>; R, <http://www.r-project.org/>; Bioconductor, <http://www.bioconductor.org/>.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturegenetics/>.

Accession codes. Raw sequences have been deposited in the EBI European Nucleotide Archive with accession number ERP000729 for worldwide rice germplasm (330 worldwide rice accessions: ERS037511–ERS037840) and with accession number ERP000106 for Chinese rice landraces (100 newly added Chinese rice landraces: ERS039279–ERS039378).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the China National Rice Research Institute for providing the rice germplasm samples. We thank S. Griffiths and G. Moore for critical reading of the manuscript. We thank Z. Zhang and E.S. Buckler for helping us use the compressed MLM and Z. Ning for assistance with sequence alignment. This work was supported by the Chinese Academy of Sciences (KSCX2-YW-N-094), the Ministry of Agriculture of China (2011ZX08001-004 and 2011ZX08009-002), the National Natural Science Foundation of China (30821004) and the Ministry of Science and Technology of China (2011CB100205) to B.H.

AUTHOR CONTRIBUTIONS

B.H. conceived of the project and its components. J.L. and B.H. contributed to the original concept of the project. W.L., Y.G., L.D., D.F., Y.L., Q.W. and Q.F. performed the genome sequencing. X.H., Q.Z., Y.Z., C.Z., K.L., L.S., T.H. and T.L. performed the genome data analyses. Y.Z., C.Z., Q.Z. and X.H. improved the imputation program for the data analyses. X.H., Q.Z. and Y.Z. developed an analytical framework for *de novo* assembly of the low-coverage sequencing data. X.W., C.L., A.W., T.Z., Y.J., G.D. and Q.Q. collected samples and performed the phenotyping. Y.Z. and X.H. performed the GWAS and statistical analyses. X.H. and B.H. analyzed all of the data together and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Weigel, D. & Mott, R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
- Zhao, K. *et al.* Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS ONE* **5**, e10780 (2010).
- Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
- Clark, R.M. Genome-wide association studies coming of age in rice. *Nat. Genet.* **42**, 926–927 (2010).
- Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
- The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
- 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Fan, C. *et al.* 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 (2006).
- Mao, H. *et al.* Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. *Proc. Natl. Acad. Sci. USA* **107**, 19579–19584 (2010).
- Saitoh, K., Onishi, K., Mikami, I., Thidar, K. & Sano, Y. Allelic diversification at the C (OsC1) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. *Genetics* **168**, 997–1007 (2004).
- Sweeney, M.T., Thomson, M.J., Pfeil, B.E. & McCouch, S. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* **18**, 283–294 (2006).
- Tian, Z. *et al.* Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. *Proc. Natl. Acad. Sci. USA* **106**, 21760–21765 (2009).
- Shomura, A. *et al.* Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
- Weng, J. *et al.* Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.* **18**, 1199–1209 (2008).
- Schneeberger, K. *et al.* Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. USA* **108**, 10249–10254 (2011).
- Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet.* **6**, e1001111 (2010).
- Yu, Y. *et al.* Independent losses of function in a polyphenol oxidase in rice: differentiation in grain discoloration between subspecies and the role of positive selection under domestication. *Plant Cell* **20**, 2946–2959 (2008).
- Ren, Z.H. *et al.* A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nat. Genet.* **37**, 1141–1146 (2005).
- Mills, R.E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).
- Sun, X. *et al.* *Xa26*, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. *Plant J.* **37**, 517–527 (2004).
- Yano, M. *et al.* Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene CONSTANS. *Plant Cell* **12**, 2473–2484 (2000).
- Dai, L. *et al.* Genomic structure and evolution of the Pi2/9 locus in wild rice species. *Theor. Appl. Genet.* **121**, 295–309 (2010).
- Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
- Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
- Li, Y., Huang, Y., Bergelson, J., Nordborg, M. & Borevitz, J.O. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 21199–21204 (2010).
- Kojima, S. *et al.* *Hd3a*, a rice ortholog of the *Arabidopsis* FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. *Plant Cell Physiol.* **43**, 1096–1105 (2002).
- Xue, W. *et al.* Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat. Genet.* **40**, 761–767 (2008).
- Zhu, B.F. *et al.* Genetic control of a transition from black to straw-white seed hull in rice domestication. *Plant Physiol.* **155**, 1301–1311 (2011).
- Huang, X. *et al.* Natural variation at the *DEP1* locus enhances grain yield in rice. *Nat. Genet.* **41**, 494–497 (2009).
- Song, X.J., Huang, W., Shi, M., Zhu, M.Z. & Lin, H.X.A. QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* **39**, 623–630 (2007).
- Cui, J. *et al.* Characterization and fine mapping of the *ibf* mutant in rice. *J. Integr. Plant Biol.* **49**, 678–685 (2007).
- Kump, K.L. *et al.* Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* **43**, 163–168 (2011).
- Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**, 2194–2202 (2009).
- Lu, T. *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-Seq. *Genome Res.* **20**, 1238–1249 (2010).
- Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
- Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).

ONLINE METHODS

Sampling and sequencing. The rice accessions were selected from a large collection of cultivated rice preserved at the China National Rice Research Institute. DNA from a single plant was used for sequencing. Seeds derived from that same plant were used for the field trials. Previously, 520 rice accessions were sequenced for the initial rice HapMap³. In this study, sampling and sequencing were improved as follows: (i) among 520 rice accessions, sequencing coverage was further scaled up for accessions with coverage lower than onefold; (ii) a further 100 Chinese *japonica* landraces were added and sequenced; and (iii) a total of 330 worldwide rice cultivars from 33 countries were sampled for sequencing, each to approximately onefold coverage. DNA isolation, library construction and Illumina sequencing were all performed as described³. In total, the genome sequences of 950 rice accessions were used in the improved rice HapMap.

Sequence alignment and genotype calling. The 73-bp paired-end reads of 620 rice landraces (with 'HP' as the ID prefix) were mapped to the rice reference genome (IRGSP Build 4.0) using the software Ssaha2 version 2.3, and the sequence reads of 330 worldwide cultivars (with 'GP' as the ID prefix) were mapped using the software Smalt version 0.2.2 (an updated version of Ssaha2) with the parameters set as "-pair 50,700" and "-mthresh 20". The detailed procedure of SNP identification was as described previously³. Genotypes of the 950 rice accessions were called at the SNP sites. For the genotype datasets of all the accessions, SNPs with more than 80% missing data and SNPs with MAF < 3% were excluded. For the genotype datasets in each subspecies, SNPs with MAF < 5% were excluded. The *k* nearest neighbor-based imputation algorithm was used to fill in missing genotypes.

Sequence assembly. For each gene, short reads of the rice accessions within the same haplotype group were combined (**Supplementary Note**). The short reads were retrieved and sorted based on their mapping information, each into a certain haplotype group of a certain gene. Once one of the paired-end reads was uniquely aligned (that is, the read was mapped to unique locations in the reference genome), we picked up both of them so as not to lose sequences in the insertion or highly polymorphic regions. For *de novo* assembly of combined sequencing data for each gene, we used a new short read assembler, Fuzzypath version 3.1, and the *kmer* value was set to 31. Variants in genic regions (including exons, introns and promoter regions) are more likely to have influence on gene coding and expression, so the assemblies of genic regions more directly facilitate follow-up study of GWAS. Hence, we focused on genic regions, and only genic regions were applied in the haplotype-based assembly. The approach can be extended to the entire genome with an appropriate LD block partition across the genome. However, non-genic regions often contain more repetitive sequences, which may make assembly in non-genic regions more difficult.

Variation detection. The resulting contigs from haplotype-based assembly were aligned back onto rice genome sequences using the software MUMmer (version 3)³⁹. Aligned contigs were picked up with a cutoff of a minimum of 92% identity over 80% of the nucleotides of a contig. The contigs, which cannot be mapped to the original gene locations, were filtered out. Sequence variations, including SNPs and indels (including highly polymorphic segments and duplication), were called using the diffseq program (using the default parameter) in the EMBOSS package⁴⁰. SNPs less than 10 bp away from each other, which were recognized to be highly polymorphic segments in the diffseq program, were retrieved and called as adjacent SNPs by a custom perl script. The error analysis of the sequence assembly and the resulting variation dataset is described in the **Supplementary Note**.

Variation annotation. SNPs and indels in coding regions were called on the basis of the gene models in the RAP-DB (release 2). The coding SNPs were then annotated to be synonymous or non-synonymous, and this information was used to calculate the nonsynonymous-to-synonymous ratio for each gene. SNPs with large-effect variations were annotated and partitioned as SNPs that introduced stop codons, SNPs that disrupt stop codons, SNPs that disrupt initiation codons or SNPs that disrupt splice sites. Indels with large-effect variations were annotated and partitioned to be frameshifting indels (indels

that produce frameshift mutation), non-frameshifting indels (indels that are a multiple of three and result in a net gain or loss of one or more amino acids) or indels that disrupt splice sites.

Phenotyping. Approximately 36 seeds for each accession from the collection of 950 accessions were germinated and planted in the experimental field (in Hangzhou, China at N 30.32°, E 120.12°) from May 2010 to October 2010. Heading date was recorded daily as the number of days from sowing to the observation of first inflorescences that emerged above the flag leaf sheath. Grain-related traits were measured in the laboratory following harvest after grains were detached from panicles and the awns were removed from the grains. We used the standard micro-Kjeldahl method to assay the crude protein content in grain⁴¹. The nitrogen conversion factor, 5.95, was used to calculate the protein content in the rice grain. Measurements of other traits were performed as described previously³.

Association analyses. GWAS of *indica*, *japonica* and the full population were conducted using their corresponding genotype datasets. In each panel, only the SNPs with MAF > 5% were used for association analyses. Both the compressed MLM model and the simple model were used in the analysis²³. In the simple model (simple linear regression model), we examined the association between the SNP genotypes and phenotypes. In the compressed MLM, the top five principal components were used as fixed effects, and the matrix of genetic similarity based on simple SNP matching coefficients (which was also used for constructing the neighbor-joining tree) was used to model the variance-covariance matrix of the random effect. The original version of the compressed MLM has some convergence problems when analyzing large samples. The parameter estimates in the covariance components were then modified to help with modeling the data using the increased sample size. Permutation tests were performed to estimate the genome-wide type I error rate (**Supplementary Note**)⁴².

Expression profile analyses. We used the Affymetrix Rice Genome Array (GPL2025) platform datasets for expression profile analyses. The microarray data files (CEL files) from each experiment were downloaded from the Gene Expression Omnibus (GEO) website. Overall, 835 experimental samples from 51 study series were used in this study. The detailed list of the microarrays and related descriptions are given in **Supplementary Table 13**. The signal intensity values for each probe were read from the CEL files using the Bioconductor packages implemented in the R language package. The data were normalized using the robust multichip average method of background correction, quantile normalization and summarization of cell signal intensities⁴³. We used transcriptional profiling of drought responsiveness in six temporal-spatial conditions (GEO series ID: GSE26280) to identify rice genes involved in drought tolerance⁴⁴. Genes that had different transcript levels between the control and drought-treated organs were identified at *P* < 0.05 using a *t*-test and had a more than twofold expression change in more than four temporal-spatial conditions. We used the root-specific transcriptional profiling of four rice varieties in response to salinity stress (GEO series ID: GSE14403) to identify rice genes involved in salt tolerance. Genes that had different transcript levels between the control and salt-treated roots were identified at *P* < 0.1 using a *t*-test and had a more than 1.4-fold expression change in more than two varieties.

Follow-up analyses for candidate genes. To seek candidate genes in the identified loci for the corresponding traits, we used public gene annotation datasets from both RAP-DB and MSU-RGAP. All the annotated genes within ~200 kb of peak SNPs were retrieved. We then used the expression dataset from the Affymetrix Rice Genome Array platform to analyze the expression pattern of each gene in the local genomic region. Among the 835 experimental samples currently available, there are 55 samples measured in the ripening seed and 84 samples measured in the panicle⁴⁵. Expression in the two tissues, the ripening seed and the panicle was calculated and compared with the average of the expression in the other tissues, as described in the **Supplementary Note**.

39. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12 (2004).

40. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
41. Tan, Y.F. *et al.* Mapping quantitative trait loci for milling quality, protein content and color characteristics of rice using a recombinant inbred line population derived from an elite rice hybrid. *Theor. Appl. Genet.* **103**, 1037–1045 (2001).
42. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
43. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
44. Wang, D. *et al.* Genome-wide temporal-spatial gene expression profiling of drought responsiveness in rice. *BMC Genomics* **12**, 149 (2011).
45. Jain, M. *et al.* F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol.* **143**, 1467–1483 (2007).

